DOCUMENT RESUME

ED 134 611 95 TM 006 019

AUTHOR	Marshall, J. Laird
TITLE	The Mean Split-Half Coefficient of Agreement and its Relation to Other Single-Administration Test Indices: A Study Based on Simulated Data. Technical Report No. 350.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Jun 76
CONTRACT	NE-C-00-3-0065
NOTE	201p.
EDRS PRICE	MF-$0.83 HC-$11.37 Plus Postage.
DESCRIPTORS	Computer Programs; *Criterion Referenced Tests; Decision Making; Mathematical Models; Norm Referenced Tests; Simulation; Standard Error of Measurement; *Statistical Analysis; *Test Reliability; True Scores
IDENTIFIERS	*Coefficient Beta; Mean Split Half Coefficient of Agreement; Test Theory

ABSTRACT

A summary is provided of the rationale for questioning the applicability of classical reliability measures to criterion referenced tests; an extension of the classical theory of true and error scores to incorporate a theory of dichotomous decisions; a presentation of the mean split-half coefficient of agreement, a single-administration test index designed to measure the internal consistency of dichotomous classifications; and information concerning the properties, under varying conditions, of this new coefficient and several other single-administration test indices, as well as their interrelationships. Simulated data were used to provide answers to questions about the behavior of coefficient beta relative to variations in score distribution, criterion level, number of examinees, number of items, and certain basic test statistics. It was determined that coefficient beta increases as the number of items increases, but in a manner different from that predicted by the Spearman-Brown prophecy formula. It was also shown that the value of the coefficient increases as the bulk of scores departs from the criterion cutoff. Relationships between coefficient beta and other test indices are presented. Most prominent among these is the indication that for unimodal score distributions, coefficient beta and Livingston's criterion referenced reliability coefficient have similar ranges of value and fluctuations over criterion level, whereas this relationship does not hold for bimodal distributions, since coefficient beta is sensitive to the mode(s) of the score distribution while Livingston's coefficient is sensitive to the test mean. (Author/RC)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED134611

T.M

TECHNICAL REPORT NO. 350

the mean split-half coefficient of agreement and its relation to other single-administration test indices: a study based on simulated data

JUNE 1976

WISCONSIN RESEARCH
AND DEVELOPMENT
CENTER FOR
COGNITIVE LEARNING

TM006 019

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Technical Report No. 350

THE MEAN SPLIT-HALF COEFFICIENT OF AGREEMENT AND ITS RELATION TO OTHER SINGLE-ADMINISTRATION TEST INDICES: A STUDY BASED ON SIMULATED DATA

by

J. Laird Marshall

Report from the Project on Conditions
of School Learning and Instructional Strategies

Thomas A. Romberg
Faculty Associate

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

June 1976

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

WISCONSIN RESEARCH AND DEVELOPMENT CENTER FOR COGNITIVE LEARNING

MISSION

The mission of the Wisconsin Research and Development Center for Cognitive Learning is to help learners develop as rapidly and effectively as possible their potential as human beings and as contributing members of society. The R&D Center is striving to fulfill this goal by

- conducting research to discover more about how children learn
- developing improved instructional strategies, processes and materials for school administrators, teachers, and children, and
- offering assistance to educators and citizens which will help transfer the outcomes of research and development into practice

PROGRAM

The activities of the Wisconsin R&D Center are organized around one unifying theme, Individually Guided Education.

FUNDING

The Wisconsin R&D Center is supported with funds from the National Institute of Education; the Bureau of Education for the Handicapped, U.S. Office of Education; and the University of Wisconsin.

ACKNOWLEDGMENTS

I would like to express my gratitude:

- to the Wisconsin Research and Development Center for Cognitive Learning, for providing me with working space, computer time, printing costs, and graphic and editorial assistance;
- to my committee as a whole, for their requirement that I restrict my scope and delineate my plans, and individually

  to Robert L. Thorndike, Chairman, for his wisdom, insightful criticism, and unwillingness to let me get away with very much;

  to Ruth Z. Gold, for her warmth, support and demands for clarity;

  to Jeremy Kilpatrick, for his patience, encouragement, support, and helpful editorial suggestions;
- to my close friend and colleague, Ed Haertel, for being, over a period of years, an excellent computer programmer, idea source, and late-night intellectual sounding board for important parts of this document;
- and to my best friend (and wife), Nancy Marshall, for being who she is; although we are both educational psychologists of sorts, my field is numbers and formulas, and hers is people and feelings; the fact that I am writing this is due in large measure to her having practiced her area of expertise on me.

TABLE OF CONTENTS

Page

Acknowledgments iv

List of Tables vii

List of Figures ix

Abstract xiii

I. Introduction 1

Behavioral Objectives, Individualized Instruction, and Mastery Learning 1
Criterion-Referenced Tests 3
Overview 5

II. Related Test Theory 9

Purpose of a Test 9
Score Distributions 11
Test Specifications and Item Selection 13
The Mathematical Model and Errors of Measurement 14
Meaning of Reliability 20

III. Coefficient Beta: The Mean Split-half Coefficient of Agreement 25

History and Rationale 25
Definitions 26
Analysis of the Coefficient 28
The Coefficient 31
Adjustment for Odd n 32
Technical Characteristics of Coefficient Beta 35
Discussion 38
Coefficient Beta and Trichotomous Data 40

IV. Other Single-Administration Coefficients 45

Livingston's Criterion-Referenced Reliability Coefficient 45
Harris's Index of Efficiency 48
The Index of Separation 50
Other Fourfold Table Test Indices 53

V. Focus of the Study, Data Generation, and Analytical Method 59

Focus of the Study 59
The Computer Program 60
The Questions and Research Methods 74

Table of Contents (cont.)

Page

VI. Results and Conclusions 77

Characteristics of Coefficient Beta 77
Characteristics of Livingston's k^2_{TX} 95
Characteristics of Harris's μ^2_c 102
Characteristics of S_c 112
Relations Among Criterion-Dependent Indices 119

VII. Summary and Suggestions for Future Research 135

Summary 135
Suggestions for Further Research 140

References 143

Appendix A: Supplementary Algebraic Derivations 149

Appendix B: Graphs of $\phi(X)$ for each Score X, for Selected Criterion Levels and Number of Items . . 155

Appendix C: Computer Program Input Parameter Distributions and Subroutines, with Notes on Calculation of Vector Components 165

Appendix D: Summaries of Stepwise Analyses of Regression . . . 169

Appendix E: A Binomial Model for Stepping Up Coefficient Beta 175

LIST OF TABLES

Table	Page

1 Errors of Measurement under Two True-Score Models. 17

2 Schema for Dual True-Score Model 20

3 A Fourfold Table for True and Observed Classifications . . 21

4 Input Parameters Used for the Study. 73

5 Values of $\phi(X)$ for $n = 20$, $c = .7$ 78

6 Ordinal Rank of Each Distribution on the Variable Indicated at Top of Column, 1 = Low, 8 = High. 84

7 Values of Spearman's Rho (Rank-Order Correlation) Between min β, $\bar{\beta}$, and Basic Test Statistics. 85

8 Values of Spearman's Rho (Rank-Order Correlation) Between max (μ_c^2), $\bar{\mu}_c^2$ and Basic Test Statistics 106

9 Extreme Fluctuations in r_{cospi} 128

LIST OF FIGURES

Figure		Page
1	$\phi_i(X)$ for a 20-Item Test; Two Criterion Levels	36
2	Two Hypothetical Score Distributions	47
3	A Cathode-Ray Tube Analogy for the Computer Program. . . .	62
4	Relationships Between Item, Examinee, and Test Characteristics; A Comparison of the Classical (A) and Computer (B) Models.	64
5	Histogram of Components of a Normally Distributed Competence Vector $(\overline{c_p})$	65
6	Histogram of Components of a Bimodal Competence Vector $(\overline{c_p})$.	66
7	Score Distribution Resulting from Parameter Set 1.	67
8	Score Distribution Resulting From Parameter Set 2.	68
9	Score Distribution Resulting from Parameter Set 3.	69
10	Score Dsitribution Resulting from Parameter Set 4.	69
11	Score Distribution Resulting from Parameter Set 5.	70
12	Score Distribution Resulting from Parameter Set 6.	70
13	Score Distribution Resulting from Parameter Set 7.	71
14	Score Distribution Resulting from Parameter Set 8.	72
15-18	Graphs of Coefficient beta against Criterion Level, with Score Distribution Relative Frequencies, for Parameter Sets 1-4	80
19-22	Graphs of Coefficient beta against Criterion Level, with Score Distribution Relative Frequencies, for Parameter Sets 5-8	81
23	Scatterplot of β for 2N (or 4N) Examinees against β for N Examinees.	86
24	Scatterplot of β (and α) for 2n Items against β (and α) for n Items.	88

List of Figures (cont.)

Figure		Page
25	Scatterplot of β for 2n Items against β for n Items, for a Normal Distribution.	90
26	Scatterplot of β for 2n Items against β for n Items, for a Uniform Distribution	92
27	Scatterplot of β for 2n Items against β for n Items, for a Bimodal Distribution	94
28-31	Graphs of k^2_{TX} against Criterion Level, with Score Distribution Relative Frequencies, for Parameter Sets 1-4	96
32-35	Graphs of k^2_{TX} against Criterion Level, with Score Distribution Relative Frequencies, for Parameter Sets 5-8	97
36	Scatterplot of k^2_{TX} for 2N (or 4N) Examinees against k^2_{TX} for N Examinees	99
37	Scatterplot of k^2_{TX} for 2n Items against k^2_{TX} for n Items.	101
38-41	Graphs of μ_c^2 against Percent Mastery, for Parameter Sets 1-4	104
42-45	Graphs of μ_c^2 against Percent Mastery, for Parameter Sets 5-8	105
46	Scatterplot of μ_c^2 for 2N (or 4N) Examinees against μ_c^2 for N Examinees.	109
47	Scatterplot of μ_c^2 for 2n Items against μ_c^2 for n Items. . .	111
48-51	Graphs of S_c against Criterion Level, with Score Distribution Relative Frequencies, for Parameter Sets 1-4	113
52-55	Graphs of S_c against Criterion Level, with Score Distribution Relative Frequencies for Parameter Sets 5-8	114

List of Figures (cont.)

Figure Page

56 Scatterplot of S_c for 2N (or 4N) Examinees Against S_c for N Examinees 116

57 Scatterplot of S_c for 2n Items Against S_c for n Items. . . 118

58 Indices vs. Criterion Level; Parameter Set 1 120

59 Indices vs. Criterion Level; Parameter Set 2 121

60 Indices vs. Criterion Level; Parameter Set 3 122

61 Indices vs. Criterion Level; Parameter Set 4 123

62 Indices vs. Criterion Level; Parameter Set 5 124

63 Indices vs. Criterion Level; Parameter Set 6 125

64 Indices vs. Criterion Level, Parameter Set 7 126

65 Indices vs. Criterion Level; Parameter Set 8 127

ABSTRACT

The report provides a summary of the rationale for questioning the applicability of classical reliability measures to criterion-referenced tests; an extension of the classical theory of true and error scores to incorporate a theory of dichotomous decisions; a presentation of the mean split-half coefficient of agreement, a single-administration test index designed to measure the internal consistency of dichotomous classifications; and information concerning the properties, under varying conditions, of this new coefficient and several other single-administration test indices, as well as their interrelationships.

Simulated data were used to provide answers to questions about the behavior of coefficient beta relative to variations in score distribution, criterion level, number of examinees, number of items, and certain basic test statistics. It was determined that coefficient beta increases as the number of items increases, but in a manner different from that predicted by the Spearman-Brown prophecy formula. It was also shown that the value of the coefficient increases as the bulk of scores departs from the criterion cutoff.

Relationships between coefficient beta and other test indices are presented. Most prominent among these is the indication that for unimodal score distributions, coefficient beta and Livingston's k^2_{TX} have similar ranges of value and fluctuations over criterion level, whereas this relationship does not hold for bimodal distributions,

since coefficient beta is sensitive to the mode(s) of the score distribution while k^2_{TX} is sensitive to the test mean.

CHAPTER I

INTRODUCTION

Behavioral Objectives, Individualized Instruction, and Mastery Learning

In the past decade, educators have given an increasing amount of attention to the related ideas of behavioral objectives, individualized instruction, and mastery learning. These ideas may be nothing more than what good teachers have been using or working toward for centuries, but it cannot be denied that formalizing and labeling them has had and will continue to have a great impact on education.

The notion that a curriculum, or at least important parts of it, can successfully be broken down into sets of behavioral objectives has been advanced by several authors (e.g., Gagné, 1965), and within the past few years there has been a progression from the theoretical to the practical, from scholarly articles to the commercial educational marketplace. Such commercially available programs as the *Wisconsin Design for Reading Skill Development* (Otto & Askov, 1974), *Developing Mathematical Processes* (Developing Mathematical Processes Staff, 1974), and *Science--A Process Approach* (American Association for the Advancement of Science Comission on Science Education, 1965) are representative of this move from theory into practice.

But educational reform has not stopped with the development of curricula based at least in part on behavioral objectives. Along with the shift toward objectives has come an increased emphasis on flexibility in instruction, to give each pupil (at least in theory) a better chance of receiving the kind of instruction that best meets his needs. One reason for such a system of individualized instruction (Klausmeier, Quilling, Sorenson, Way, & Glasrud, 1971) is that individuals in a given group do not all learn a given set of materials at the same rate or by the same methods, a fact which has been all too painfully obvious to generations of teachers faced with pupils on one end of the ability spectrum who exhibited boredom and pupils on the other end who felt frustrated when they have used a pace and form of presentation appropriate for some pupils in the middle.

A system of behavioral objectives and individualized instruction, however, offers hope: the objectives allow the teacher to concentrate on a discrete block of material, and individualization improves the chances that a given student will spend neither more nor less time on the material than is needed. This, of course, raises the question, "How much time is 'enough'?" Although this question is so open-ended as to have frustrated many theoreticians and researchers, a good bit has been written on the topic, which has come to be known as the "mastery learning" issue. While much of the current interest in mastery learning was given impetus by an article by Bloom (1968), the underlying philosophy has profited from contributions of many writers (e.g., Carroll, 1963).

One can easily discuss mastery learning in a theoretical way, but to make the concept operational in a classroom means defining mastery for a given behavioral objective, and this in turn necessitates describing the method by which mastery is to be assessed. This description does not usually present too great a difficulty; if a behavioral objective is explicitly stated, it is generally possible to explicate how mastery can be assessed. Evans (1968) claims, however, that the behavioral objectives are less important operationally than the assessment instrument. He maintains that the posttest, not the list of behavioral objectives, is the ultimate operational measure of what a teacher is trying to teach. While mastery may sometimes have to be assessed by somewhat uncommon methods, this report will only concern itself with the familiar paper-and-pencil test format.

Criterion-Referenced Tests

There are several kinds of instruments whose stated purpose is to assess mastery. They differ in the number of objectives involved, the number of items per objective, nomenclature, the meaning of criterion, and the interpretation given to the test results.

Some tests measure only one objective (DMP Staff, 1974); others encompass several objectives. Of these, some test each objective with a single test item (Gessel, 1972) while others require more than one.

There are several names given by various writers to these assessment instruments: mastery test, objectives-based test, objective-referenced measure, domain-referenced test, and criterion-referenced test. This last term, introduced over a decade ago (Glaser, 1963) has gained perhaps the widest currency. Such widespread use has also resulted in widespread abuse, since this single term is employed to cover a range of test types and interpretations. Recognizing this problem, Donlon (1974) and Millman (1974) have offered schemata for labeling various kinds of criterion-referenced tests.

In addition, some authors disagree on the meaning of the word criterion. Some writers (e.g., Nitko, 1971) maintain that criterion means some observable standard of performance; others (e.g., Harris & Stewart, 1971) define it as a specified percentage of correct responses on test items. Some writers indicate that interpretation of the test results should take into account how many items were responded to correctly or how far from the criterion the examinee's score lies, whereas others maintain that the sole matter of importance is whether mastery was attained. At an even more basic level, there are writers (Simon, 1969) who argue that there is no such thing as a criterion-referenced test separate from a more traditional norm-referenced test; rather, the interpretation one puts on the score (absolute number rather than relative ranking) is the basis for the distinction.

Any of these viewpoints may have merit; however, for the purpose of this report, a criterion-referenced test (CRT) is defined as a test that measures performance on a single behavioral objective, that has several items drawn from a well-defined universe, and whose results yield a dichotomous mastery/nonmastery decision with reference to a predetermined criterion level expressed as a percentage of items answered correctly. As such, it comes closest to Roudabush's (1974) category of a pseudo-continuous measure of a dichotomous true score. It also seems to fall into Millman's (1974) category of a CRDAD, or criterion-referenced differential assessment device, although this writer does not agree with all the nuances of implication of the CRDAD classification. Some of these areas of disagreement will be discussed in the next chapter.

It will also be shown in the next chapter that a CRT, as defined above, differs from the more familiar norm-referenced test in several fundamental aspects: purpose, test specifications, desired score distributions, method of reporting scores, and meaning of reliability, among others. Thus the two kinds of tests are quite different and, although they share some properties, one kind is not, for example, a special instance or a generalization of the other.

Overview

This report deals with CRTs as previously defined, and its major focus is on the notion of CRT reliability. Because the purposes, construction, application, and psychometric theory of CRTs are considered by many to differ from those of norm-referenced tests (NRTs), serious questions have been raised in recent years as to whether classical reliability measures ought to be applied to CRTs.

In Chapter II, several of these questions are raised and investigated, and an attempt is made to show that classical reliability indices are not meaningful for at least one important aspect of CRTs. An extension of the classical mathematical model that incorporates this aspect of CRTs is suggested and a definition of CRT reliability is presented. Also suggested is a set of criteria for a CRT reliability index.

Chapter III is an exposition of coefficient beta, the mean split-half coefficient of agreement (Marshall & Haertel, 1975), a recently developed single-administration CRT reliability coefficient. This new coefficient is based on the theory presented in Chapter II and meets the criteria suggested therein.

In Chapter IV, a few other CRT indices that have been presented in the recent literature, including those of Livingston (1972a) and Harris (1972a), are discussed with emphasis on how well they meet the criteria suggested in Chapter II. In addition, other indices used in this study are defined.

Chapter V presents the questions investigated in this study concerning properties of coefficient beta and its relations to other test indices. The statistical methodology utilized in answering these questions is described, as is the computer program used to generate the simulated data for the study.

Chapter VI presents the results of these investigations and draws a number of conclusions, and Chapter VII offers a summary and suggests areas for future research.

The purpose of this report is to provide the educational measurement community with:

1. a brief summary of the rationale for questioning the applicability of classical reliability measures to CRTs.
2. an extension of the classical theory of true and error scores to incorporate a theory of dichotomous decisions.
3. a detailed presentation of the mean split-half coefficient of agreement, a new single-administration test index designed to measure the internal consistency of dichotomous classifications.
4. systematic data concerning the properties, under varying conditions, of this new coefficient and several other single-administration test indices, as well as their interrelationships.

In summary, this report offers the rationale, the exposition, the characteristics, and the relationship to other test indices of a new coefficient designed to measure the dichotomous decision-making reliability of CRTs.

CHAPTER II

RELATED TEST THEORY

It is appropriate to examine how a criterion-referenced test (CRT) (as defined in Chapter I) differs from a norm-referenced test (NRT). A number of authors have discussed aspects of the subject using various definitions of a CRT (Brennan, 1974; Glaser, 1963; Glaser & Cox, 1968; Hambleton & Novick, 1973; Millman, 1974; Popham & Husek, 1969). In this chapter, certain parts of classical test theory will be discussed briefly and extended to incorporate a proposed theory for CRTs. The discussion will include the interrelated topics of the purpose of a test, score distributions, test specifications and item selection, the underlying mathematical model and errors of measurement, an extension of the mathematical model, and the meaning of reliability. Since this chapter is not a treatise on measurement theory as such, the discussion will not cover all areas in detail but will instead focus on those points that bear on the arguments developed here.

Purpose of a Test

The fundamental purpose of an NRT is to differentiate among individuals by assigning to each examinee a number, or estimated true score, in reference to the norms of the population for which the test is designed. One's score on an NRT indicates a level of achievement that is given meaning by comparison with the group at large; it is a measure of relative standing within the group that can be communicated via grade equivalents, standard deviations above and below the mean, stanines, centiles, "grading on a curve," etc. (It is true that NRTs can be used for dichotomous decisions--a person may be selected for admission to a training program, for example, according to whether he scores above a certain cutoff point--but this cutoff score is chosen in reference to the performance of other candidates and thus is different from the criterion cutoff score on a CRT.)

Not too many years ago a commissioner of education, in a public policy address, indicated his hope that within a certain period of time everyone would be reading "up to or above grade level." When one considers that grade level is another term for mean, this comment reduces to a proposal that everyone should be at or above average. Although the statement is humanistically generous, it is statistically self-contradictory.

Given a well-defined behavioral objective, however, one could correctly make a statement about everyone's performing at or above a certain criterion level. One could measure this performance with a CRT as defined in Chapter I. The purpose of such a CRT is not to rank individuals or to report scores in reference to a norm, but rather to enable one to make a dichotomous decision based on whether a given pupil is performing on a given behavioral objective at or above a certain predetermined level (as defined by a certain score or percent correct on the CRT.) Thus the purpose of a CRT is different from that of an NRT: a CRT provides data from which to make a decision on an absolute, not a relative, standard (see also Glaser, 1963.)

Score Distributions

Since, as stated earlier, the purpose of an NRT is to discriminate among examinees, one would naturally hope for a fairly even score distribution with a wide spread of scores, so as to allow efficient discrimination. Thus, in theory, the optimum total score distribution for an NRT would have some shape within the range of normal (with large standard deviation), platykurtic, rectangular, or slightly bimodal (with modes at the extremes). Total score distributions of this sort usually enhance test reliabilities since they produce moderately high total-score variance. In practice, and consistent with most theories of traits within a population, a large-variance normal or a symmetrical, somewhat platykurtic, distribution often obtains.

However, the assumption of a normal or a platykurtic distribution for competence on a given behavioral objective[1] is clearly contradictory to the reason for and purpose of instruction. The reason for giving instruction toward an objective is that students have not mastered it; one assumes that before instruction, student proficiencies are massed for the most part at the lower end of the spectrum. In teaching, one hopes that all students will master the objective. With individualized instruction some students may take a good deal longer than others, but ultimately the purpose of this instruction is to ensure that the mass of student proficiencies shifts to the upper end of the scale. In neither case is a normal distribution implied.

---

[1] Here, and elsewhere in this paper, attention is restricted to a certain limited type of behavioral objective--one that is quite specific and narrow in scope, usually from the cognitive domain, and measurable by a test of several items.

To quote Bloom (1968),

> If we are effective in our instruction, the distribution of achievement should be very different from the normal curve. In fact, we may even insist that our educational efforts have been unsuccessful to the extent to which our distribution of achievement approximates the normal distribution [p. 3].

With a CRT, moreover, used to make a dichotomous decision with respect to a predetermined criterion, the desired discrimination is not among individuals but rather between two mastery groups--those students who have met the objective and those who have not (Glaser & Cox, 1968.) Hence the desired score distribution is one that is rather sharply bimodal, with one mode well below and the other mode rather above the cutoff point (Roudabush, 1974). Research by Blatchford (1970) shows that these bimodal distributions do indeed occur in classroom testing. He comments, "In a diagnostic test, as [an example of] a criterion-referenced test, there is no evidence of a normal distribution [p. 43]."

For a given administration of a CRT, particularly before or immediately after instruction, it is even plausible (and quite acceptable) for the set of scores for one of these mastery groups to be empty or very nearly so, producing small variance and hence distorted estimates of reliability by traditional means (Stanley, 1971). Much has been made of this point in the literature (e.g., Popham & Husek, 1969). Thus the need arises for a new definition of CRT reliability, so that a test's reliability estimate is not adversely affected by a score distribution with small variance. It will also become evident, in the next few sections, that there are additional difficulties in applying a traditional reliability estimate to a CRT.

Test Specifications and Item Selection

Although a detailed discussion of the mechanics of test specifications and item selection is not within the scope of this report, certain facets of this practical topic should be mentioned.

In the construction of traditional tests, the domain of the test is often defined in relatively loose terms such as first-year biology, or reading comprehension, or mathematical aptitude. First a test blueprint is prepared indicating in broad outline the processes and topics to be covered. Then items are selected; if they fit the test blueprint and if they fall within the purview of the subject matter, they are fair game for inclusion in the initial version of the test. Whether they are included in the final version depends on performance on them in the test tryouts (and whether the items taken as a whole still fit the blueprint). Decisions regarding an item's suitability for the final version are usually made in terms of its difficulty and either the item-test correlation (Davis, 1952) or an analogous statistic (e.g., Baker, 1965).

A CRT, on the other hand, has a decidedly narrower focus, delineated by the behavioral objective, and thus the items admissible for inclusion in a test tryout must meet far more restricted specifications. Some writers have claimed that traditional methods of item selection are therefore inappropriate and have offered alternative methods based on from one to three test administrations (Brennan, 1974; Brennan & Stolurow, 1971; Cox & Vargas, 1966; Darlington & Bishop, 1966; Ivens, 1970; Kosecoff & Klein, 1974; Millman, 1974; Popham, 1971; Popham & Husek, 1969; Wedman, 1973).

Definitive answers to the question of how best to choose items for CRTs are still being sought, but the work of these authors implies that traditional methods probably are not the solution.

The Mathematical Model and Errors of Measurement

In classical test theory, the usual mathematical model defines an examinee's observed score on a test as comprising two components: true score and error. This model is usually expressed by an equation equivalent to $X_p = T_p + E_p$, where p is the subscript for persons. Here E_p is the error of measurement, the amount by which a person's obtained score (X_p) differs from his true score (T_p), which in turn is the score that would have been obtained with a (purely theoretical) perfect measuring instrument or would have been derived from an infinite number of administrations of the test or parallel versions of it. This kind of model has been thoroughly discussed in the literature (e.g., Lord & Novick, 1968) and will not be detailed here.

It is important to note that several assumptions are associated with this mathematical model and hence with its derived results. Three of these assumptions (Lord & Novick, 1968, p. 56) are basic to the definition of classical reliability and are mentioned here, since they are scrutinized later. These assumptions are that (1) true score and error have zero covariance, (2) the expected value of error over persons is zero, and (3) errors on parallel measurements have zero covariance.

In classical theory, the basic question asked is, What is the examinee's true score? True score is considered a continuous variable and is expressed on a scale that is usually considered to be interval, if not ratio. Observed score, expressed on the same scale, is usually a polytomous rather than a continuous variable, but only because of the nature and limits of the measuring instrument, which ordinarily produces scores with integral values. Hence error, like true score, is continuous; in absolute value, it is expressed on a scale that is ratio.

This continuous true-score model serves nicely when the purpose of the test is to determine as precisely as possible what one's true score is and to report that estimated true score on a polytomous scale. But that is not the fundamental purpose of a CRT as defined in this paper. Rather, the basic question asked by a CRT is, "Is the examinee's true score great enough to allow him to be placed in the 'mastery' classification?" Although the continuous true score is used as a first step in answering this question, the final answer, or decision, is dichotomous and is reported on a scale that is ordinal but not interval.

These facts suggest an alternative model--one of dichotomous true and observed scores, with score in this sense meaning decision. This model has been labeled Platonic (Sutcliffe, 1965) and is well summarized by Lord and Novick (1968, pp. 39-44). Although the equation $X = T + E$ is unchanged, elsewhere this model differs markedly from the classical model presented above. First, true score and observed score, being dichotomous variables, are expressed not on an interval scale but on an ordinal scale, as is error, which is a trichotomous variable (or dichotomous in absolute value). Second, Klein and Cleary (1967) have shown, among other things, that with the Platonic true-score model, the covariance of true and error scores is generally negative and is zero only under extraordinary circumstances. They also have shown that the expected value of Platonic error scores is not likely to be zero, and that errors on parallel tests cannot be expected to have zero covariance. All three of these findings violate the assumptions upon which the derivation of classical test reliability rests. (Since covariance is a statistic designed for interval data, one could question why it has been computed for a model whose data are measured on an ordinal scale. One could similarly question the computation of a variance or a correlation and thus the applicability of a classical reliability coefficient for dichotomous data.)

The meaning of measurement error is also different for the two models. In classical theory, it is the examinee's true score and hence the size of the error present in the obtained score that are the psychometrician's subjects of interest. The obtained score, if there is error, can vary from the true score by a lot or by a little, and it makes a difference to the psychometrician which of these cases holds. In the Platonic model, however, there is only one kind of measurement error--incorrect categorization. There is no great or small associated with it; the psychometrician is concerned with the existence, not the size, of error. This view has been stated succinctly by Cronbach and Gleser (1965) as follows: "a test designed to be maximally efficient for a particular decision will freely allow errors to enter if they are irrelevant to that decision [p. 137]." Others (Hambleton & Novick, 1973) have recognized, even without accepting the Platonic model, that there is only one kind of measurement error for a CRT.

TABLE 1

ERRORS OF MEASUREMENT UNDER TWO TRUE-SCORE MODELS

Student	Classical Theory			Platonic Theory		
	X	T	E	X	T	E
A	15	9.4	5.6	0	0	0
B	16	20.0	-4.0	1	1	0
C	15	19.5	-4.5	0	1	-1
D	16	10.8	5.2	1	0	1
E	15	16.2	-1.2	0	1	-1
F	16	15.2	.8	1	0	1

Moreover, classical and Platonic measurement error need not correspond for a given set of data. Consider the hypothetical data in Table 1 for a 20-item CRT with a mastery criterion of 80%, yielding 16 as the cutoff score. Of students A through F, all of whom have an obtained score of 15 or 16, students A, B, C, and D have the largest classical measurement error and students C, D, E, and F have the largest (only) Platonic measurement error. Likewise, the students with the smallest classical measurement error are not necessarily those with a Platonic measurement error of 0.

The table shows that, given the distributions of observed and true scores under the two models, there need not be a high correlation between classical and Platonic measurement error, particularly when observed scores are very near the cutoff score. These data, of course, have been chosen to illustrate a point, and as observed scores begin to move away from the cutoff, the correlation between the two kinds of measurement error will increase. However, if the scores move farther from the cutoff, the correlation will decrease. In any case, classical and Platonic measurement errors are different things, and the theory developed for one kind of error need not apply to the other.

This fact raises the question of which theoretical model is appropriate, or preferable, for CRTs. There are arguments for both models. Those supporting the classical model argue that even if a CRT is designed to make a dichotomous decision, its initial results (observed score) are reported on a polytomous scale. It is also felt that a dichotomous decision "often hides the true level of student performance [Klein & Kosecoff, 1973, p.9]." Supporters of this model believe that it is just not realistic to claim that a person's true score on a behavioral objective is an all-or-nothing entity.

Primary among the arguments supporting the Platonic model is the belief that when a test is used to make a dichotomous decision--"go on" or "don't go on" to the next behavioral objective--the size of the obtained score is immaterial except as it results in a mastery or nonmastery classification. It is felt that this dichotomous score is the only one that need be reported; further subdivisions of the obtained score have no practical value. "Such gradations in reporting [scores] are only a function of the alternative courses of action available to the individual after the measurements have been made [Popham & Husek, 1969, p.8]."

It appears that one must choose between the model that is consistent with continuous true and error scores and the model that incorporates dichotomous decision and error scores. The need to choose between these two models can be avoided (and, it is claimed here, should be avoided) by broadening one's view of the meaning of true and observed scores. The contention here is that classical true-score theory is appropriate when the basic purpose of a test is to estimate the true score. But when the test has a different basic purpose, such as to determine a dichotomous classification, then the examinee has not one true score but two, existing simultaneously: a true score that is involved with the primal measuring process and another that has to do with the decision or basic question to be answered concerning the individual and thus with the practical results of that measurement. It can even be said that there are as many different sets of "true scores" as there are alternate score-reporting schemes.

The assertion in this report is that a CRT as defined here involves two different facets of true score--positional and operational. The first facet deals with the position of one's test score in relation to the test scores of others; the second facet deals with the operational effects of the test score on the examinee alone. Classical NRT theory concerns itself only with the former and for good reason. When the end result of the testing process is to associate the examinee with a number (when the test's basic question is what his true score is), then the positional and operational facets are indistinguishable. But when the end result of the testing process is to make a dichotomous decision (when the basic question is whether the examinee merits a certain classification) and the outcome of that decision has an immediate and differentiating effect on the student's next educational activity, then the difference between these positional and operational facets emerges.

The dual true-score model for CRTs is summarized in Table 2.

TABLE 2

SCHEMA FOR DUAL TRUE-SCORE MODEL

Facet	Basic Question to be Answered	Equation	Scale of Answer
Positional	What is the true score?	$X = T + E$	Continuous
Operational	Is the true score high enough to merit "mastery" classification?	$D = C + M$*	Dichotomous

*D = observed classification (Decision), C = true Classification, M = Misclassification (error)

Meaning of Reliability

Classical reliability can be defined as the squared correlation between observed and true scores (Lord & Novick, 1968, p. 61). This statistic is equal to the ratio of true-score variance to observed-score variance if the conditions noted at the beginning of the previous section are assumed. The classical true-score model, presented here as the positional facet of the dual true-score model, is consistent with those assumptions and therefore with these definitions of reliability. However, the Platonic model, or operational facet of a CRT, is not consistent with those assumptions (Klein & Cleary, 1967), and hence the classical notion of reliability cannot apply whenever the reliability of a test has to do with the consistency of decision making, i.e., whenever the basic measurement question is to be answered dichotomously.

What then should be the meaning of the operational reliability of a CRT? For the positional facet, a test is reliable insofar as an examinee receives the same relative ranking on two sets of data (and in the case of parallel tests, the same score); for the operational facet, a CRT should be reliable insofar as an examinee receives the same classification on both sets of data. Put differently, positional reliability is concerned with the accuracy of assigning (polytomous) numbers to examinees; operational reliability (henceforth called CRT reliability) must necessarily be concerned with the accuracy of placement in one of two categories.

Consider the theoretical fourfold contingency table given in Table 3. Classical reliability is defined in terms of a mathematical relationship between true and observed scores. It would be natural to begin to investigate CRT reliability in the same terms. With reference to Table 3, one approach would be to consider the squared correlation between true and observed classifications, $\rho^2(C,D)$. Since the variables are dichotomous, this would imply the use of the squared

TABLE 3

A FOURFOLD TABLE FOR TRUE AND OBSERVED CLASSIFICATIONS

		Observed Classification (D)	
		+	-
True Classification (C)	+	a	b
	-	c	d
			N

phi coefficient if the dichotomy is a true one. However, the dual true-score model presented previously and the arbitrariness of the mastery cutoff score of a CRT suggest that true classification is an artificial rather than a real dichotomy, and hence that the phi coefficient is not the appropriate statistic. (Nonetheless, the phi coefficient is calculated from a different fourfold table in the investigation presented in Chapters V and VI.)

If the dichotomy is artificial, then the tetrachoric correlation coefficient is the appropriate statistic and would yield a formula in a, b, c, and d. (See Table 3). The objection to the cosine-pi estimate of this statistic is that if either a or d is 0, then the correlation is -1 even though it may be near 1 when a or d is merely close to 0. (Nonetheless, the cosine-pi estimate of the tetrachoric correlation coefficient is also calculated from a different fourfold table in the study presented later.)

Another approach to the mathematical relationship between C and D is the variance-ratio approach. As pointed out earlier, one cannot assume zero covariance between true classification and misclassification (error). When this assumption is rejected, a true-classification variance/obtained-classification variance ratio of

$$\frac{(a+b)(c+d)}{(a+c)(b+d)} = \frac{\pi(1-\pi)}{p(1-p)}$$

is obtained, where π is the true proportion of mastery classification and p is the obtained proportion of mastery classifications. But this statistic is unsatisfactory for at least two reasons. First, if $a = d$ or $b = c$, then $r = 1$ no matter what numbers are in the other two cells;

second, if $.5 < \pi < p$ or $p < \pi < .5$, $r > 1$, which is clearly not acceptable.

So it appears that for the true (C) and observed (D) classifications in Table 3, neither the correlation approach nor the ratio of variances approach yields a satisfactory coefficient. Thus there must be some other mathematical relationship between C and D that affords a meaningful CRT reliability index. One such relationship, which follows directly from the notion of CRT reliability as consistency of classification, is the proportion of classifications that are correct classifications, $\frac{a + d}{N}$. Since a and d are unknown, it would seem that a meaningful CRT reliability coefficient would be a statistic that estimates, or perhaps is a lower bound for, this quantity. Furthermore, any such CRT reliability coefficient should have, so far as possible, the following characteristics:

1. It should be associated with the notion of consistency or accuracy of (dichotomous) classification; hence the more the scores depart from the cutoff point, the higher the CRT reliability index should be, since such a departure most clearly represents a separation between the mastery and nonmastery categories.

2. It should be, at least in some respects, variance-free, so that it will not vanish when total score variance approaches 0.

3. It should avoid any reliance on classical measurement error concepts, since they are not necessarily relevant to a test whose purpose is to make a dichotomous decision.

4. It should be a function of the criterion level, since the criterion level is an integral part of the CRT as defined in this report.
5. It should if possible have a familiar range of values, most probably [0,1], for ease of interpretation.

A coefficient that incorporates these features will be presented in the next chapter.

CHAPTER III

COEFFICIENT BETA: THE MEAN SPLIT-HALF COEFFICIENT OF AGREEMENT

History and Rationale

Some decades ago, the single-administration reliability, or internal consistency, of a test was estimated by calculating the Pearson product-moment correlation between two halves of a test, adjusted by the Spearman-Brown prophecy formula. Later, other split-half formulas were introduced (Flanagan, 1937; Rulon, 1939). But there were objections to the split-half method, since the particular test split chosen (usually odd-numbered items versus even-numbered items) was not necessarily representative, and a misleading reliability estimate could result. Other methods were proposed and proved useful (Hoyt, 1941; Kuder & Richardson, 1937). Then Cronbach (1951) showed that his coefficient alpha was not only a generalization of the Kuder-Richardson formula 20 and equal to Hoyt's internal consistency measure, but was also equal to the mean of all possible split-half reliability coefficients (but not equal to the mean of all possible stepped-up split-half correlation coefficients, see Novick & Lewis, 1967). Thus was established the basis for estimating internal consistency for a test designed to rank-order the examinees.

However, when the purpose of a test is to dichotomize rather than rank-order, the procedure to follow is not so clear-cut (Popham & Husek, 1969). Several authors (Berger, 1970; Carver, 1970; Goodman & Kruskal, 1954; Hambleton & Novick, 1973; Millman, 1974) have suggested using a simple coefficient for such test reliability, but only in the dual-administration sense. This index, given various names and symbolic labels by various authors, will here be called the coefficient of agreement and, for the sake of simplicity, labeled P. According to Goodman and Kruskal (1959), this measure of association was reported as early as 1884, although it was not used for test reliability. The suggested index is simply the proportion of individuals who are classified the same way (mastery/mastery or nonmastery/nonmastery) by two sets of data --test-retest or parallel forms. The coefficient has not been adapted to the split-half single-administration case, perhaps for the same reasons as those cited previously for the classical split-half coefficients.

However, Cronbach's (1951) finding suggests a lead: one can consider an index that would be equal to the mean of all possible split-half coefficients of agreement. To extend the analogue with Cronbach's coefficient alpha, this index will be labeled coefficient beta (β).

Definitions

Let

N = the number of people taking the test

n = the number of items in the test

X_p = the pth person's total score, $p = 1, \ldots, N$

c = the criterion level, expressed as a fraction ($0 < c \leq 1$)

k = the smallest integer $\geq \frac{cn}{2}$, and hence the minimum number of items in a half-test[2] that must be answered correctly to receive a "mastery" classification on that half-test

---

[2] For now, only tests with an even number of items are considered. Tests with an odd number of items are dealt with later in the chapter.

X_{1p}, X_{2p} = the pth person's scores for the two half-tests, and hence $X_{1p} + X_{2p} = X_p$.

There are $\binom{n}{n/2} = \nu$ possible test splits for an n-item test if one considers each half to be labeled (i.e., for a two-item test the split 1 / 2 is different from the split 2 / 1.) For each pair of split-halves, construct a fourfold mastery (+) / nonmastery (-) contingency table:

	+	-	
+	A	B	
-	C	D	
			N

and define

$$P = \frac{A + D}{N} \quad .$$

Then β is the mean of P taken over all ν possible split-halves (s):

$$\beta = \frac{1}{\nu} \sum_{s=1}^{\nu} P_s$$

$$= \frac{1}{\nu} \sum_{s} \frac{A_s + D_s}{N} \quad .$$

But $A_s + D_s$ is the number of consistent classifications (among the N persons) on test split s, and hence can be written

$$A_s + D_s = \sum_{p=1}^{N} \delta_{ps} \ ,$$

where $\delta_{ps} = 1$ or 0 as the pth person's classifications are consistent or inconsistent, respectively, on test split s. Thus β can be written

$$\beta = \frac{1}{\nu} \sum_{s=1}^{\nu} \left(\frac{\sum_{p=1}^{N} \delta_{ps}}{N} \right)$$

$$= \frac{\sum_{s} \sum_{p} \delta_{ps}}{\nu N}$$

$$= \frac{1}{N} \sum_{p} \frac{\sum_{s} \delta_{ps}}{\nu} \qquad [1]$$

Thus β is also the mean (over persons) proportion (over test splits) of consistent classifications.

Analysis of the Coefficient

For any given test, the set of possible scores for an individual is $\{0, 1, \ldots, n\}$. For computational purposes this is partitioned into five subsets, one or more of which may be empty for a particular n and k:

$S_1 = \{0, \ldots, k-1\}$

$S_2 = \{k, \ldots, 2k-2\}$

$S_3 = \{2k-1\}$

$S_4 = \{2k, \ldots, \frac{n}{2} + k-1\}$

$S_5 = \{\frac{n}{2} + k, \ldots, n\}$.

(Note that $k = 1$ implies $S_2 = \{\ \}$, and $k = \frac{n}{2}$ implies $S_4 = \{\ \}$.)

Then consider scores in each of the five subsets:

1. For $X_p \in S_1$, $X_p < k$. Thus mastery on a half-test cannot be obtained no matter how the test is split, since both X_{1p} and X_{2p} must necessarily be less than k. Hence all persons with $X_p \in S_1$ will contribute to D, as defined in the contingency table above, for all ν test splits.

2. For $X_p \in S_2$, $k \leq X_p \leq 2k-2$. Here some splits will contribute to B or C (for example, $X_p = k+1$; $X_{1p} = k$, $X_{2p} = 1$) and some will contribute to D (for example, $X_p = 2k-2$; $X_{1p} = X_{2p} = k-1$). The obvious question "Which splits?" becomes a problem of combinatorics. Since only A and D enter into Equation 1, one need not be concerned with contributions to B and C. (These contributions will be equally divided among B and C because of the symmetry implied in "labeling" the halves of the test.)

The question then reduces to "For a score of $X_p \in S_2$, how many D-categorizations will result?" A D-categorization will happen when neither half-test is mastered and thus both X_{1p}, $X_{2p} \leq k-1$.

Define $\overrightarrow{X_{1p}}$ and $\overrightarrow{X_{2p}}$ as vectors of 0's and 1's, indicating incorrect and correct responses, respectively, to items on each half-test. If one vector has k-1 1's, the other has $X_p - (k-1)$ 1's. Moreover, since $X_p \in S_2$ and hence $X_p \leq 2k-2$, it follows that $X_p - (k-1) \leq k-1$. Thus one is interested only in those pairs of vectors in which the number of 1's in each is between these two limits, namely $X_p - (k-1) \leq$ both X_{1p}, $X_{2p} \leq k-1$. Moreover, since in the total score there are X_p 1's, there are $n - X_p$ 0's. In the half-score, if there are j 1's, there are $\frac{n}{2} - j$ 0's. Thus, for $X_p \in S_2$, we can pick pairs of vectors that will yield D-categorizations in

$$\sum_{j=X-(k-1)}^{k-1} \binom{X}{j} \binom{n-X}{\frac{n}{2}-j}$$

ways.

3. For $X_p \epsilon S_3$, $X_p = 2k-1$. Thus the most "balanced" split will yield k 1's in one vector and k-1 1's in the other, indicating mastery in the first case and nonmastery in the second. Other, less "balanced" splits will yield more extreme allocations of 1's, resulting in the same mastery/nonmastery classifications. Thus, for all $X_p \epsilon S_3$, no split contributes to A or D.

4. For $X_p \epsilon S_4$, $2k \leq X_p \leq \frac{n}{2} + k-1$. This case is similar to that of S_2. Some splits will contribute to B or C (for example, $X_p = 2k$; $X_{1p} = k+1$, $X_{2p} = k-1$) and some will contribute to A (for example, $X_p = 2k$; $X_{1p} = X_{2p} = k$). Since $X_p \geq 2k$, it cannot be that both X_{1p}, $X_{2p} < k$, and hence there are no contributions to D. Again we ignore the contributions to B and C, but should focus attention instead on the contributions to A.

In this case, one needs to count those vectors where both half-tests are mastered, i.e., where both X_{1p}, $X_{2p} \geq k$. If one half-test vector contains k 1's, the other contains X_p-k 1's. But $X_p \epsilon S_4$ implies $X_p \geq 2k$, which implies $k \leq X_p -k$. Thus one is interested only in those half-test vectors such that k $\leq$ both X_{1p}, $X_{2p} \leq X_p -k$. By using reasoning identical to that for S_2, the total number of splits that will contribute to A for $X_p \epsilon S_4$ is

$$\sum_{j=k}^{X-k} \binom{X}{j}\binom{n-X}{\frac{n}{2}-j} .$$

5. For $X_p \epsilon S_5$, $X_p \geq n/2 + k$. This says that half the items plus at least another k items are answered correctly, and thus both X_{1p}, $X_{2p} \geq k$ no matter how the test is split. Hence all v splits contribute to A.

The coefficient

The above analysis yields an equation for β, the mean split-half coefficient of agreement. For X_p in each of the five subsets, define the following functions $\phi_i(X)$, $i = 1, \ldots, 5$:

1. for $0 \leq X \leq k-1$ $\qquad \phi_1(X) = 1$

2. $k \leq X \leq 2k-2$ $\qquad \phi_2(X) = \sum_{j=X-(k-1)}^{k-1} \binom{X}{j}\binom{n-X}{\frac{n}{2}-j} \Big/ \binom{n}{\frac{n}{2}}$

3. $X = 2k - 1$ $\qquad \phi_3(X) = 0$

4. $2k \leq X \leq n/2 + k-1$ $\qquad \phi_4(X) = \sum_{j=k}^{X-k} \binom{X}{j}\binom{n-X}{\frac{n}{2}-j} \Big/ \binom{n}{\frac{n}{2}}$

5. $n/2 + k \leq X \leq n$ $\qquad \phi_5(X) = 1$

Here, $\phi_i(X)$ is the proportion of splits that contribute to A or D for a given score X.

Then Equation 1 can be rewritten

$$\beta = \frac{1}{N} \sum_{p=1}^{N} \phi_i(X_p), \qquad [2]$$

where the index i depends on the value of X_p. Hence β has range [0,1]; it is 0 when all $X_p \in S_3$, and 1 when all $X_p \in S_1 \cup S_5$.

Although Equation 2 sums up the analysis rather simply, it is inefficient for computing purposes. A more efficient method involves generating a frequency distribution of total scores and computing $\phi_i(X)$ only once for each possible value. In general, let f_x be the frequency of score x, $x = 0, \ldots n$, $\sum_{X=0}^{n} f_x = N$. Then

$$\beta = \frac{1}{N} \sum_{X=0}^{n} f_x \, \phi_i(X),$$

where again the index i depends on the value of X.

More explicitly, since for some values of X, $\phi_i(X) = 0$ or 1,

$$\beta = \frac{1}{N}\left(\sum_{X=0}^{k-1} f_x + \sum_{X=k}^{2k-2} f_x \, \phi_2(X) + \sum_{X=2k}^{\frac{n}{2}+k-1} f_x \, \phi_4(X) + \sum_{X=\frac{n}{2}+k}^{n} f_x \right)$$

$$= \frac{1}{N}\left(\sum_{X=0}^{k-1} f_x + \sum_{X=k}^{2k-2} f_x \, \phi_x(X-[k-1],\ k-1) + \sum_{X=2k}^{\frac{n}{2}+k-1} f_x \, \phi_x(k,\ X-k) + \sum_{X=\frac{n}{2}+k}^{n} f_x \right) \qquad [3]$$

where

$$\phi_x(a,b) = \sum_{j=a}^{b} \frac{\binom{X}{j}\binom{n-X}{\frac{n}{2}-j}}{\binom{n}{\frac{n}{2}}} .$$

Adjustment for odd n

For an odd number of items, a test split is defined as resulting when one item is deleted and the remaining items are divided into two sets, each containing $\frac{n-1}{2}$ items. In this case, k is the smallest integer greater than or equal to $\frac{c(n-1)}{2}$. The item to be deleted may be chosen in n ways, each yielding a distinct set of n-1 items to be split. Hence there are $n\binom{n-1}{(n-1)/2}$ possible split halves, if one again considers each half to be labeled.

For person p, with total score X_p, the response vector $\overrightarrow{X_p}$ contains X_p 1's and $(n-X_p)$ 0's. Thus, for person p, X_p of the n possible choices of the item to be deleted will result in a set of n-1 items containing (X_p-1) 1's, and $n-X_p$ choices will result in a set containing X_p 1's. Thus the contribution to β for person p, rather than $\phi_i(X_p)$, will be $\frac{X_p}{n}\phi_i(X_p-1) + \frac{n-X_p}{n}\phi_i(X_p)$ and hence, taking the mean over persons,

$$\beta = \frac{1}{nN}\sum_{p=1}^{N}\left[X_p\,\phi_i(X_p-1) + (n-X_p)\,\phi_i(X_p)\right].$$

As before, it is necessary to compute $\phi_i(X)$ only once for each possible value of X.

Also as before, the computation is more efficient if we utilize the frequency distribution of total scores. Recall that for a score of X_p on n (odd) items, for $n-X_p$ choices of the item to be deleted the total score on n-1 items will remain at X_p, and for X_p choices the total score on n-1 items will be reduced to X_p-I. The effect is that of a transformation, $\xrightarrow{t}$, on the set of total scores. In symbols,

$X \xrightarrow{t} X$ in $\frac{n-X}{n}$ of the cases;

$X \xrightarrow{t} X-1$ in $\frac{X}{n}$ of the cases, and hence

$X+1 \xrightarrow{t} X$ in $\frac{X+1}{n}$ of the cases.

Thus a total score of X is arrived at with frequency

$g_X = \frac{n-X}{n} f_X + \frac{X+1}{n} f_{X+1}$. (Note that, since $f_{n+1} = 0$,

$g_n = \frac{n-n}{n} f_n + \frac{n+1}{n} f_{n+1} = 0$, and therefore $\sum_{X=0}^{n} g_X = \sum_{X=0}^{n-1} g_X$.

Furthermore, it is easily shown (see Appendix A) that $\sum_{X=0}^{n-1} g_X = \sum_{X=0}^{n} f_X$.) Thus taking the mean over the transformed frequency distribution of total scores, coefficient beta is

$$\beta = \frac{1}{N} \sum_{X=0}^{n-1} g_X \cdot \phi_i(X)$$

$$= \frac{1}{N} \left[\sum_{X=0}^{n-1} \frac{n-X}{n} f_X + \frac{X+1}{n} f_{X+1} \right] \phi_i(X),$$

where once again the index i depends on the value of X. Thus, in practice, the computation of β is identical for the cases of even and odd n, except that in the latter case one first performs an additional step, replacing f_X by $\frac{(n-X)f_X + (X+1)f_{X+1}}{n}$ for $X = 0, 1, \ldots, n-1$ and then using $n-1$ in place of n in the computations of k and $\phi_i(X)$.

Technical Characteristics of Coefficient Beta

Although coefficient beta is defined solely on the basis of fourfold contingency tables, its computational formula (Equation 3) is a function of the score distribution as well as of the number of items and the criterion level. Since these latter two parameters are (or should be) known before a test is administered, the value of β for a particular tryout results from the frequency distribution of total scores. The same is true of values of Harris's μ_c^2 and the criterion-referenced index of separation (S_c), which are discussed in the next chapter. Like S_c but unlike μ_c^2, β is the mean of its additive parts. That is, given β' for a set of scores of N-1 examinees, if the score of an Nth examinee were to be added to the set, a new β could be calculated from

$$\beta = \frac{1}{N}\left[(N-1)\ \beta' + \phi_i(X_N)\right],$$

since from Equation 2,

$$(N-1)\ \beta' = \sum_{p=1}^{N-1} \phi_i(X_p).$$

A similar argument holds for the addition of a set of scores.

Since this additivity is a property of coefficient beta, one can investigate the relative contribution of the pth person's score to the value of the coefficient, given the number of items and the criterion level, merely be determining $\phi_i(X_p)$. For illustration, Figure 1 shows these relative contributions for a 20-item test with criterion levels of 70% and 80%. Additional graphs, covering a range of numbers of items and criterion levels, can be found in Appendix B.

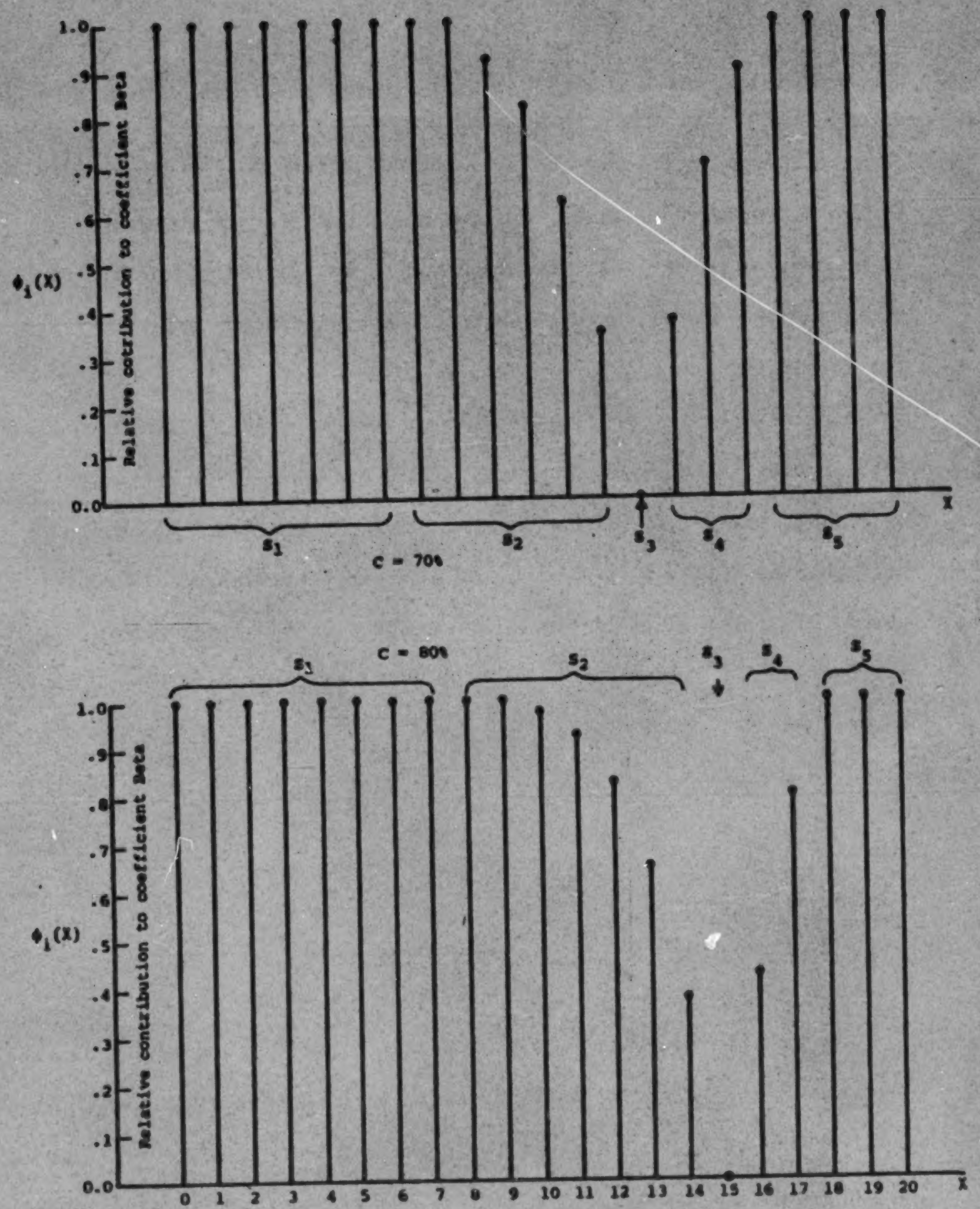


Figure 1. $\phi_i(X)$ for a 20-item test; two criterion levels

It is apparent from Figure 1 (as well as from the analysis of the coefficient presented earlier in this chapter) that as scores approach the integer immediately below the cutoff, they contribute successively less to the value of β; at the score 2k-1 (with k as defined earlier), the contribution is zero. This is to be expected since the score 2k-1 composes the subset S_3 as defined earlier and $\phi_3(X) = 0$.

Figure 1 might be misleading in the sense that, in these two examples, the point 2k-1 is one less than cn, the product of the criterion level and the number of items, and hence is one less than the test's cutoff score. One might therefore ask why 2k-1, and not 2k, is the score with a zero contribution to coefficient beta. It should be pointed out, however, that this relation does not always hold. On a 12-item test with criterion level of 75%, for example, the points 2k-1 and cn are both 9. In general, if cn is an odd integer, 2k-1 = cn; if cn is even, 2k-1 = cn-1. If cn is not an integer, 2k-1 can be greater than cn (e.g. if n = 16 and c = 80%, then cn = 12.8 and 2k-1 = 13.). In general, depending on the values of c and n, 2k-1 falls somewhere in the half-open interval [cn-1, cn+1).

Even though 2k might at first glance seem to be a more appropriate candidate than 2k-1 for the score with zero contribution to coefficient beta, 2k falls in the interval [cn, cn+2), and therefore, in a mathematical expectation sense, is not as good an approximation to cn as 2k-1.

Discussion

Although attention in this dissertation has been given to criterion-referenced tests, it should be pointed out that coefficient beta is applicable whenever reliability is viewed as consistency of classification or consistency of decision-making based on scores from a measuring instrument, provided that the classification decision is based on some sort of cutoff point expressible as a percent of items responded to in a certain manner.

Second, and consistent with the notion of accuracy of categorization from the results of a limited number of items, it should be noted that coefficient beta increases as the number of items on a test increases, as shown in a later chapter. The degree to which this increase follows the Spearman-Brown prophecy formula is discussed in Chapter VI.

Third, one should also note that if examinees respond randomly to the items on a test, the resulting coefficient beta is not zero, as might be expected with a traditional reliability measure. In fact, depending on the values of c, n, and N and on the number of options per item (assuming a multiple-choice test), coefficient beta would probably take on a rather high value, possibly even 1. From the standpoint of traditional test theory, this is disconcerting. Yet it is understandable, from the CRT standpoint, if one recalls that coefficient beta is designed to measure the operational reliability of a CRT. If all examinees respond randomly to a test, it is a clear indication that they are about as far from mastery as is possible. The high value of coefficient beta would indicate that the test is classifying most of them as such, and reliably so. Nonetheless, a test constructor might want additional test tryout information before passing judgment on the instrument's reliability, as in the construction of an NRT.

Fourth, it is appropriate to see how coefficient beta measures up to the criteria for a CRT reliability coefficient that were set forth at the end of the last chapter.

1. Coefficient beta is based on the notion of accurate placement in categories. It turns out that beta does attain its highest values when the test scores depart from the cutoff; however, these scores need not be at the extremes for beta to take on its highest values. For example, on a 20-item test with a criterion level of 70% (yielding a cutoff score of 14), $\beta = 1$ if all scores are in $\{0, \ldots, 6\} \cup \{17, \ldots, 20\}$. As the total scores pile up near the cutoff, the value of β decreases.

2. Coefficient beta is variance-free in the respect deemed most important by critics of a variance-dependent CRT reliability coefficient: it can take on any value from 0 to 1 even though the total score variance is 0, depending on the relative values of the cutoff score and the (single-membered) set of test scores. The coefficient is, however, variance-dependent in other respects. As the variance approaches its maximum, β approaches 1. This relation is reassuring since maximum variance on an n-item test occurs only when scores are equally divided between 0 and n, which scores indicate the clearest possible separation of examinees into two classifications. Furthermore, if $\beta = 0$, then the variance is zero. These relations are easily summarized: if the variance is high, coefficient beta is high; if the variance is low, there is no restriction (within its range) on coefficient beta.

3. Coefficient beta is not based on traditional measurement error concepts. Since it is built around the theory of dichotomous categorizations and Platonic true scores, the Platonic notion of misclassification is the only measurement error involved.

4. Coefficient beta is an algebraic function of the criterion level (and other parameters).

5. Coefficient beta has a range of [0,1], although values near 0 occur only under highly improbable conditions.

Coefficient beta and trichotomous data

The authors of some commercial instructional programs, such as Developing Mathematical Processes (DMP Resource Manual, Topics 1-40, 1974), contend that mastery/nonmastery alone is not a sufficient categorization of test results, and that more valuable information and more appropriate teacher options become available if the test result data are trichotomized into classifications such as "mastery," "progress," and "nonmastery." Coefficient beta, as outlined above, is clearly not sensitive to such a trichotomization scheme.

The trichotomous coefficient of agreement in such a situation would be equal to

$$P = \frac{A+E+I}{N} \quad ,$$

based on the following table, in which +, *, and - stand for the three categorizations:

$$\begin{array}{c|c|c|c|c} & + & * & - & \\ \hline + & A & B & C & \\ \hline * & D & E & F & \\ \hline - & G & H & I & \\ \hline & & & & N \end{array}$$

A coefficient analogous to β and applicable to this situation should be equal to $\frac{1}{\nu} \sum_{s=1}^{\nu} \frac{A_s + E_s + I_s}{N}$, or the mean split-half trichotomous coefficient of agreement.

Such a coefficient can be derived, although the derivation is not presented here. The analysis of this coefficient, although more complex in places, is essentially parallel to the analysis of coefficient beta presented earlier. Instead of partitioning the set $\{0, \ldots, n\}$ into five subsets, one partitions it into seven. Recall that for coefficient beta, k is the minimum number of items on a half-test that must be answered correctly for a mastery classification. If, for trichotomized data, one in addition lets ℓ be the minimum number of items on the half-test that must be answered correctly for the middle classification, then the seven subsets of $\{0, \ldots, n\}$, together with their corresponding values of $\phi_i(X)$, $i = 1, \ldots, 7$, are

$$S_1 = \{0, \ldots, \ell-1\} \qquad \phi_1(X) = 1$$

$$S_2 = \{\ell, \ldots, 2\ell-2\} \qquad \phi_2(X) = \sum_{j=X-(\ell-1)}^{\ell-1} \binom{X}{j} \binom{n-X}{\frac{n}{2}-j} \Big/ \binom{n}{\frac{n}{2}}$$

$$S_3 = \{2\ell-1\} \qquad \phi_3(X) = 0$$

$$S_4 = \{2\ell,\ldots,2k-2\} \qquad \phi_4(X) = \sum_{j=u_1}^{u_2} \binom{X}{j}\binom{n-X}{\frac{n}{2}-j} \Big/ \binom{n}{\frac{n}{2}}$$

$$S_5 = \{2k-1\} \qquad \phi_5(X) = 0$$

$$S_6 = \{2k,\ldots,\tfrac{n}{2}+k-1\} \qquad \phi_6(X) = \sum_{j=k}^{X-k} \binom{X}{j}\binom{n-X}{\frac{n}{2}-j} \Big/ \binom{n}{\frac{n}{2}}$$

$$S_7 = \{\tfrac{n}{2}+k,\ldots,n\} \qquad \phi_7(X) = 1$$

where $0 < \ell < k \leq \frac{n}{2}$,

$u_1 = \max(\ell,\ X-[k-1])$,

and $u_2 = \min(k-1,\ X-\ell)$.

Note that $\ell = 1$ implies $S_2 = \{\ \}$ and $k = \frac{n}{2}$ implies $S_6 = \{\ \}$.

As before, the computation is made more efficient by utilizing the frequency distribution of total scores, and hence a formula for β_3, the mean split-half trichotomous coefficient of agreement, is

$$\beta_3 = \frac{1}{N}\sum_{X=0}^{n} f_x\ \phi_x(X).$$

Since $\phi_i(X)$ is 0 or 1 in four of the seven cases, this can be more explicitly rewritten as

$$\beta_3 = \frac{1}{N}\left[\sum_{X=0}^{\ell-1} f_x + \sum_{X=\ell}^{2\ell-2} f_x\ \phi_x(x-[\ell-1],\ \ell-1) + \sum_{X=2\ell}^{2k-2} f_x\ \phi_x(u_1,u_2) + \sum_{X=2k}^{\frac{n}{2}+k-1} f_x\ \phi_x(k,\ X-k) + \sum_{X=\frac{n}{2}+k}^{n} f_x\right],$$

where $$\phi_x(a,b) = \sum_{j=a}^{b} \frac{\binom{X}{j}\binom{n-X}{\frac{n}{2}-j}}{\binom{n}{\frac{n}{2}}}$$

and u_1 and u_2 are as above.

The trichotomous coefficient requires the same adjustments for an odd number of items as does the dichotomous coefficient, except that $n-1$ is used in calculating ℓ as well as k and $\phi_1(X)$.

Note that if the test is multiple choice, the lower of the two criterion levels should not be set near the percent of items that should be answered correctly due to chance, as this would result in unreliable classification decisions between the lower two categories. In this case, if there are a significant number of nonmasters in the population, the value of β_3 would tend to be rather low, as would be expected.

CHAPTER IV

OTHER SINGLE-ADMINISTRATION COEFFICIENTS

Several authors have recently devised or resurrected indices dealing either directly or peripherally with CRT reliability. Some indices are based on one administration of a test (Harris, 1972a; Livingston, 1972a; Marshall, 1973), some on two administrations (Berger, 1970; Carver, 1970; Hambleton & Novick, 1973; Ivens, 1970; Millman, 1974; Ozenne, 1971; Swaminathan, Hambleton, & Algina, 1974), and some on three administrations (Brennan, 1974). This report is concerned solely with single-administration indices.

The two single-administration coefficients that have received the widest attendtion are k^2_{TX} (Livingston, 1972a) and μ^2_c (Harris, 1972a). A third measure is the index of separation of test scores (Marshall, 1973). These and three other coefficients are presented in this chapter. Since the relation of each of these indices to coefficient beta is detailed in a subsequent chapter, their rationale is discussed briefly here, as is their degree of adherence to the criteria given at the close of Chapter II.

Livingston's Criterion-Referenced Reliability Coefficient

Livingston's coefficient, k^2_{TX}, is widely known and the most discussed coefficient in the recent literature. It stems from an interesting application of classical reliability theory, and departs therefrom only in the notion of mean square deviation. Instead of using variance as the mean square deviation from the mean of scores, Livingston substitutes for it a quantity equal to the mean square deviation from the cutoff point. The assumption is that the deviation of a person's score from the cutoff, not the deviation from the mean, is of interest in a CRT. The rest of Livingston's careful algebraic development parallels that of classical theory, and the resulting k^2_{TX} is related algebraically to classical quantities:

$$k^2_{TX} = \frac{r\sigma^2 + (\overline{X}-C)^2}{\sigma^2 + (\overline{X}-C)^2} \qquad [4]$$

where

r = classical internal consistency reliability

σ^2 = variance of total scores

$\overline{X}$ = mean of total scores

C = criterion cutoff point (not necessarily an integer).

As can be seen from Equation 4 (and as pointed out by Livingston), $k^2_{TX} \geq r$, and k^2_{TX} approaches r as $\overline{X}$ approaches C.

This coefficient has been the subject of much criticism, comment, and rebuttal (Hambleton & Novick, 1973; Harris, 1972b; Hsu, 1971; Livingston, 1972b, 1972c; Marshall, 1973; Ozenne, 1971; Raju, 1973; Shavelson, Block, & Ravitch, 1972). Summaries of the arguments can be found in the references by Brennan (1974), Rim (1974), and Wedman (1973), and are not presented here. In this section, k^2_{TX} is analyzed with respect to the criteria for a CRT reliability index set forth at the end of Chapter II.

1. It is not the distances of the scores themselves from the criterion cutoff that contribute to high values of k^2_{TX}, but rather the distance of the mean of scores from the cutoff, as Equation 4 shows. This fact is of no consequence when the score distribution is unimodal and generally symmetric, since under these conditions the mode and mean will tend to coincide. But when the distribution is bimodal, which is desirable for a CRT, then this fact becomes important in interpreting k^2_{TX}; it is particularly important when the mean falls about halfway between the two modes. Consider the earlier example of a 20-item test with a cutoff of 14. Suppose the data from two samples, A and B, form "inverted triangular" distributions with different means, as shown in Figure 2. If the classical test reliability is .80 in both cases, k^2_{TX} = .91 for sample A and .80 for sample B, even though sample B seems to show a clearer separation between nonmasters and masters, since there are fewer scores at or near the cutoff. (Coefficient beta would have values of .72 and .88 for samples A and B, respectively.)

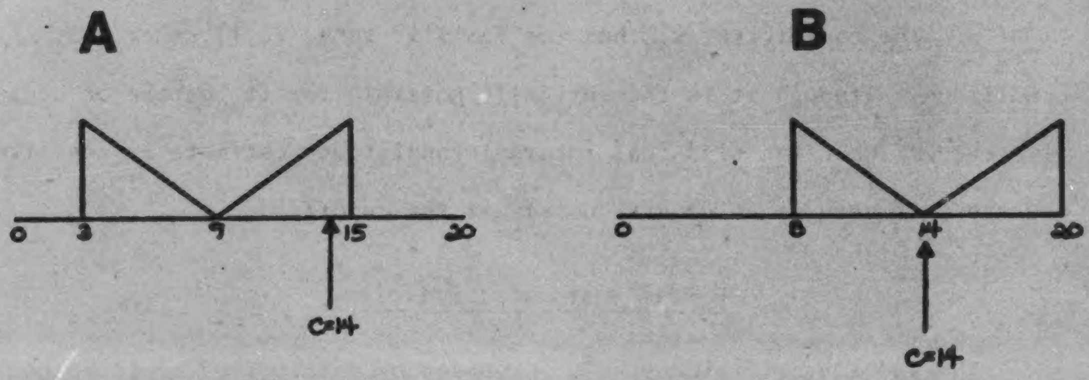


Figure 2. Two hypothetical score distributions.

2. The coefficient k^2_{TX} is not variance-free, as is evident from Equation 4; it is dependent on total-score variance, as are classical coefficients, although in a different way. When total score variance is zero, $k^2_{TX} = 1$ (see Equation 4), unless $\overline{X} = C$. Thus the coefficient does not vanish when the variance approaches zero, but instead it tends toward a unique value. When the variance approaches its maximum, k^2_{TX} again approaches 1 because the traditional reliability coefficient also approaches 1 under these conditions and $k^2_{TX} \geq r$. Under less extreme conditions total score variance has varying effects on k^2_{TX}.

3. The coefficient k^2_{TX} is based on classical error of measurement. In fact, as Harris (1972b) points out in criticizing the coefficient, the standard error of measurement is the same in Livingston's framework as it is in the classical framework, even though the value of Livingston's reliability coefficient is normally higher than that of a classical coefficient.

4. As Equation 4 shows, k^2_{TX} is an algebraic function of the criterion level (and other parameters).

5. The coefficient k^2_{TX} has the familiar range [0,1] under most conditions, although it is theoretically possible for it to take on negative values, when the classical internal consistency estimate is negative and the test mean is at or very nearly at the cutoff.

Harris's Index of Efficiency

The index of efficiency, μ^2_c, proposed by Harris (1972a) is intended "to examine how well the test sorts defined samples of students into categories and possibly to measure its efficiency in this sense [p. 4.]".

It has been interpreted as the squared correlation between test score and a 0 / 1 dummy variable representing the nonmastery/mastery classifications. Harris also points out that μ_c^2 can be conceived of as the ratio of true-score variance to observed-score variance if true score is defined for the subjects in each of the two groups as the group mean.

The computational formula is

$$\mu_c^2 = \frac{SS_b}{SS_b + SS_w}, \qquad [5]$$

where the terms in the ratio represent the between-group and within-group sums of squares for the groups resulting from the dichotomous classification.

The index is analyzed as follows with respect to the CRT reliability criteria.

1. The index of efficiency has highest values when the total score distribution is sharply bimodal with a mode on either side of the cutoff, but these modes need not be far from the cutoff. For illustration, given the 20-item test with $C = 14$, $\mu_c^2 = 1$ if all scores are either 0 or 20, which is reassuring. But μ_c^2 is also 1 if all scores are 13 or 14; a perfect μ_c^2 occurs even though all mastery/nonmastery classifications could be reversed with a change of only one point in each person's total score. (Coefficient beta would have a very low value under these conditions--less than 0.20 if the scores are more or less evenly divided--and Livingston's k_{TX}^2 would be no greater than $.5r + .5$, where r is a classical reliability coefficient.)

2. The index of efficiency is variance-dependent, but in a somewhat different way than a classical coefficient is. As Equation 5 indicates, μ_c^2 is undefined when total-score variance is zero; and when total-score variance is at its maximum, $\mu_c^2 = 1$. But μ_c^2 can also be high even though the variance is small (but not zero). Given a 20-item test with a cutoff of 14, $\mu_c^2 = 0$, if all examinees score 14 or 15; if one examinee scores 13 and the rest score 14, $\mu_c^2 = 1$.

3. Except for the true-variance/total-variance ratio interpretation mentioned earlier, the index of efficiency is not based on traditional measurement error concepts. (An example of the index's departure from traditional measurement error concepts was given under point 1.)

4. Although not explicitly part of the computing formula, the criterion level is nonetheless implicit in the calculation of μ_c^2 since it is the basis for defining the two groups into which the examinees are sorted and for which the sums of squares are calculated.

5. The index of efficiency has the familiar [0,1] range. It is 0 when all examinees are classified the same way (provided variance is not 0); it is 1 when there are two groups and each within-group variance is 0 (see Equation 5).

The Index of Separation

The index of separation of total scores (S) is designed to measure the degree to which the set of total scores on an n-item test approaches the set {0,n}. It is based on the assumption that the population taking a CRT is in fact the union of two subpopulations, either of which may be empty: one knowledgeable, and hence with expected test score $E(X) = n$; the other not knowledgeable, with $E(X) = 0$, either when the test is free-response or when the scores are corrected for guessing. The formula for this index is

$$S = 1 - \frac{4}{nN} \sum_{p} \left(X_p - \frac{1}{n} X_p^2 \right), \qquad [6]$$

where n and N are the numbers of items and persons, respectively, and X_p is the pth person's total score.

An alternative formulation for S is

$$S = \frac{4}{n^2 N} \sum_{p} \left(\frac{n}{2} - X_p \right)^2 .$$

If this is rewritten as

$$S = \frac{\frac{1}{N} \sum_{p} \left(X_p - \frac{n}{2} \right)^2}{\frac{n^2}{4}} ,$$

it follows that S can be interpreted as the ratio of A to B, where A is the mean square deviation of the X_p from n/2 and B is the maximum possible mean square deviation from n/2 (and hence the maximum possible variance for a test of n items.)

The index can be analyzed according to the CRT reliability criteria as follows:

1. The index of separation has maximum values insofar as scores depart from n/2 rather than from the cutoff. Thus S is a score distribution index and is not criterion-dependent; this is also clear from Equation 6.

2. The index of separation is algebraically related to total score variance by the formula

$$S = 1 - 4\left(\overline{p}\,\overline{q} - \frac{\sigma_x^2}{n^2}\right) ,$$

where $\overline{p}$ is the mean item difficulty (i.e., $\frac{\Sigma X}{nN}$) and $\overline{q} = 1 - \overline{p}$. Nonetheless, S is variance-free in the same important respect as coefficient beta is: it can take on its full range of values even though the total score variance is zero. Also like coefficient beta, S = 1 when variance is at its maximum, and S = 0 implies zero variance (when the set of total scores is {n/2}).

3. The index of separation is independent of classical measurement error concepts.

4. The index of separation is not a function of the criterion level; it is a function of the frequency distribution of total scores alone. Its value for a given score distribution is therefore invariant under changes in the criterion level. Thus it is a score distribution index and not a CRT index.

5. The index of separation has range [0,1]. It is 0 when all scores are n/2, and 1 when all scores are 0 or n.

Since the index of separation fails to satisfy criteria 1 and 4, it may be helpful to introduce a related index that satisfies these criteria. Such an index, the criterion-referenced index of separation (S_c), is formulated as follows:

$$S_c = \frac{1}{N}\left[\sum_{X \le C} f_x\left(\frac{C-X}{C}\right)^2 + \sum_{X > C} f_x\left(\frac{X-C}{n-C}\right)^2\right] \qquad [7]$$

where f_x is the frequency of score X in the score distribution.

Appendix A demonstrates that $S_c = S$ if $C = \pi/2$, and thus S_c is a generalization of S. The criterion-referenced index of separation meets all five CRT index criteria. Thus coefficient beta will be compared with it as well as with the Livingston and Harris coefficients in Chapter VI.

Other Fourfold Table Test Indices

In the analyses reported in Chapter VI, reference is made to two other indices besides those CRT coefficients discussed thus far. In this section, these other indices are described.

First, consider again the definition of the elements of the mastery (+) / nonmastery (-) contingency table:

	+	-
+	A	B
-	C	D
		N

and recall that coefficient beta is equal to the mean of all possible split-half coefficients of agreement, where the coefficient of agreement, P, is

$$P = \frac{A + D}{N} \quad .$$

The cosine-pi estimate A correlation statistic, appropriate when the two (inherently continuous) underlying variables have been artificially dichotomized, is the cosine-pi estimate (r_{cospi}) of the tetrachoric correlation coefficient (r_{tet}). A computing formula is

$$r_{cospi} = \cos \frac{\pi}{1 + \sqrt{AD/BC}} , \qquad [8]$$

where the angle is expressed in radians and the symbols A, B, C and D refer to the entries in the contingency table above. This formula yields a good estimate of r_{tet} only when the marginal frequencies of the contingency table do not depart markedly from $\frac{1}{2}$ N (Guilford, 1965).

The phi coefficient Another index is the phi coefficient (r_ϕ). Its formula is

$$r_\phi = \frac{AD - BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}} , \qquad [9]$$

where A, B, C, and D are defined as before. The phi coefficient is a special case of the Pearson product-moment correlation that is calculated on two inherently dichotomous variables.

Normally, the computation of these coefficients requires two sets of data (resulting from two administrations of a test). However, in the course of the computer calculation of coefficient beta, a "grand" fourfold table with entries equal to the means of the results of all possible split-half categorizations is easily constructed. It follows from the analysis of the derivation of coefficient beta given in Chapter III, and from Equation 3 in particular, that the entries in the cells of this "grand" fourfold table are:

$$A^* = \sum_{X=2k}^{\frac{n}{2}+k-1} f_X \left[\sum_{j=k}^{X-k} \frac{\binom{X}{j}\binom{n-X}{\frac{n}{2}-j}}{\binom{n}{\frac{n}{2}}} \right] + \sum_{X=\frac{n}{2}+k}^{n} f_X$$

$$D^* = \sum_{X=0}^{k-1} f_X + \sum_{X=k}^{2k-2} f_X \left[\sum_{j=X-(k-1)}^{k-1} \frac{\binom{X}{j}\binom{n-X}{\frac{n}{2}-j}}{\binom{n}{\frac{n}{2}}} \right]$$

$$B^* = C^* = \frac{1}{2}(N - A^* - D^*) \quad .$$

In this study, the cosine-pi estimate and the phi coefficient are calculated from this "grand" table, and under these conditions they can be construed as single-administration indices. Note, for example, that the r_ϕ thus calculated is not equal to the mean of all possible split-half phi coefficients--the computer program was not designed to do the calculations required--but rather is a single coefficient calculated from a table resulting from all possible split-half nonmastery/mastery categorizations.

Coefficient kappa Millman (1974) and Swaminathan et al. (1974) have proposed that coefficient kappa (Cohen, 1960), an index originally developed for nominal data, rather than the coefficient of agreement, is the appropriate index to use for dual-administration CRT reliability. The computing formula for κ is

$$\kappa = \frac{p_o - p_c}{1 - p_c} ,$$

where, in the case of dichotomous categorization for each administration,

$$p_o = \frac{A}{N} + \frac{D}{N},$$

the observed proportion of like categorizations (i.e., the coefficient of agreement), and

$$p_c = \left(\frac{A+B}{N}\right)\left(\frac{A+C}{N}\right) + \left(\frac{B+D}{N}\right)\left(\frac{C+D}{N}\right),$$

the "expected" (by chance) proportion of like categorizations (i.e., the sum of products of marginal proportions, as in a chi-square test of association).

The advantage claimed for coefficient kappa is that it removes from the final coefficient that proportion of agreement due to chance, that is, the expected proportions in the population. It seems unclear, however, what interpretation should be given to the notion of population proportion. In the case of an attribute with truly nominal values, say eye color, it makes sense to talk of the proportion of the population with hazel eyes (given, of course, a suitable measurement process for identifying "hazel"). But for such an ephemeral attribute as degree of mastery of a given behavioral objective, where fifteen minutes of instruction may well change a person from the non-mastery category to the mastery category, the "expected proportion" of the population in one category is not so clear.

Since coefficient kappa is a dual-administration index, it is not within the scope of this study. It would be interesting, however, to consider a single-administration coefficient equal to the mean of all possible split-half kappa coefficients. Unfortunately, the algebra involved is forbidding.

However, rather than take the mean of all possible split-half kappa coefficients, one can treat coefficient kappa in the same way as the cosine-pi estimate and the phi coefficient; namely, one can calculate the kappa coefficient from the "grand" fourfold table, which gives the means of all possible split-half categorizations. Recall from the derivation of coefficient beta, that cells B and C in the "grand" fourfold table are equal because of the symmetry implied in labeling the halves of the test. To indicate this, let

$$B = C = E,$$

and let the superscript (*) denote a coefficient calculated from the "grand" table. Then

$$\kappa^* = \frac{P_o - P_c}{1-P_c} = \frac{\frac{A}{N} + \frac{D}{N} - \left[\frac{(A+E)\ (A+E)}{N^2} + \frac{(D+E)\,(D+E)}{N^2}\right]}{1 - \left[\frac{(A+E)\,(A+E)}{N^2} + \frac{(D+E)\,(D+E)}{N^2}\right]}$$

With a little algebra (see Appendix A) this can be simplified to

$$\kappa^* = \frac{AD - E^2}{(A+E)\,(D+E)}$$

Note, however, that under these same conditions of B = C = E, the phi coefficient (from Equation 9) is

$$r_\phi^* = \frac{AD - E^2}{\sqrt{(A+E)(A+E)(D+E)(D+E)}}$$

$$= \frac{AD - E^2}{(A+E)(D+E)} \quad . \qquad [10]$$

Therefore, when B = C, the kappa coefficient (κ^*) is identical with the phi coefficient (r_ϕ^*). Moreover if one denotes the mean split-half kappa coefficient by $\bar{\kappa}$, κ^* seems to be a lower bound to $\bar{\kappa}$. (This hypothesis results from some limited empirical paper-and-pencil research by this author, and is based on a comparison of the two quantities calculated from a few manufactured score distributions and some item-by-pupil response matrices associated with each distribution. All examples supported this tentative result.) Moreover, the difference between the two expressions is usually slight, generally not more than .05. For example, a hypothetical 4-item test with a mastery criterion level of 100%, and ten examinees with a total score vector of (0,1,1,1,2,3,3,3,4,4), yields

$$r_\phi^* = \kappa^* = .282.$$

Four different item-by-pupil response matrices yield values of $\bar{\kappa}$ ranging from .287 to .362. Thus, in interpreting the results concerning the phi coefficient presented in Chapter VI, one should bear in mind that r_ϕ^* appears to be a (generally close) lower bound to $\bar{\kappa}$.

It is also of interest to note that $1 - r_\phi^*$ or $1 - \kappa^*$ is identical with $\hat{I}$ calculated from a fourfold table with equal off-diagonal cells, where $\hat{I}$ is the estimate of the index of inconsistency used for binomial data by the Bureau of the Census, as reported by Cochran (1968, p. 663).

CHAPTER V

FOCUS OF THE STUDY, DATA GENERATION, AND ANALYTICAL METHOD

This chapter details the questions to be investigated, the means by which the data were obtained, and the methodology of statistical analysis.

Focus of the Study

The study was designed to answer the following four questions, which are presented in detail later in the chapter:

1. What are the characteristics of coefficient beta relative to variations in score distribution, criterion level, number of examinees, number of items, and certain basic test statistics?
2. What are the characteristics of the three other criterion-dependent test indices defined in Chapter IV?
3. Are there predictable relationships between coefficient beta and any or all of these three indices?
4. Are there predictable relationships between coefficient beta and other fourfold contingency table indices?

Large amounts of systematic data are needed to obtain satisfactory answers to these questions. Prohibitive expenditures of resources and inordinate cooperation from schools would be required to collect such data empirically, and hence the data were simulated by computer.

The Computer Program

A computer program was designed by the investigator and written by a colleague to generate the data for the study. The purpose and design of the program were threefold: (1) to simulate the results of the test-taking process by generating item-by-person response matrices of 0's and 1's; (2) to allow for systematic control of the generation of these matrices by providing great flexibility in the definition of input parameters, to be discussed later; and (3) to create graphic aids and to calculate various statistics, including those used in this study, from each simulated response matrix.

The Basic Equation

The first step in using the computer program is to define the input parameters, discussed in the next section. Then the program generates a response matrix of 0's and 1's according to the equation

$$r_{ipt} = g_i^2 [(c_p + e'_{pt}) - (1 - d_i)] + (1 - g_i^2)[e''_{ip} + e_{ipt}] \qquad [11]$$

where

r_{ipt} is the response to the ith item by the pth person on the tth trial (or replication) of the test;

g_i is the "goodness" of an item, akin to item-test correlation, with range [0,1];

c_p is the "competence" of the person on the behavioral objective being measured, with range [0,1];

d_i is the "facility" and therefore $1-d_i$ is the intrinsic difficulty of an item, with range [0,1];

and the e's are normally distributed random components each with an expected value of 0, but whose variance may be specified. The first is a persons-by-trials component: persons feel different from day to day and would react to tests differently as a result. The second is a (generally larger) items-by-persons component: it is not realistic to assume that a given item will have the same difficulty, relative to other items, for each person. The third is a catch-all, undefined component that varies over items, persons, and trials, and may be thought of as related to errors of measurement.

The response is counted as correct or incorrect, and thus the element in the response matrix is 1 or 0 as $r_{ipt} \geq 0$ or $r_{ipt} < 0$, respectively.

Note that when an item is perfectly "good," $g_i = 1$, and when the persons-by-trials error component is ignored, Equation 11 reduces to

$$r_{ipt} = c_p - (1 - d_i) \quad ,$$

implying that the response is recorded as correct when the person is at least as competent as the item is difficult. Further, a perfectly "bad" item would be one with $g_i = 0$; in this case the basic equation [11] reduces to

$$r_{ipt} = e''_{ip} + e_{ipt} \quad ,$$

implying that the correctness of the response is due completely to random factors. Note further that for a perfectly good item the effect of the item-by-person error vanishes; this effect does appear when the item is not perfectly good. These values of item goodness are limits rather than realities, of course, and thus the values of g_i actually used in the investigation were between these extremes.

In order to clarify how the basic equation functions, consider the cathode-ray tube analogy shown in Figure 3. Think of the value of $c_p + e'_{pt}$ as an emission point on a cathode, and consider the value of $1 - d_i$ as a hole in a grid. Then $c_p + e'_{pt} - (1 - d_i)$ could be

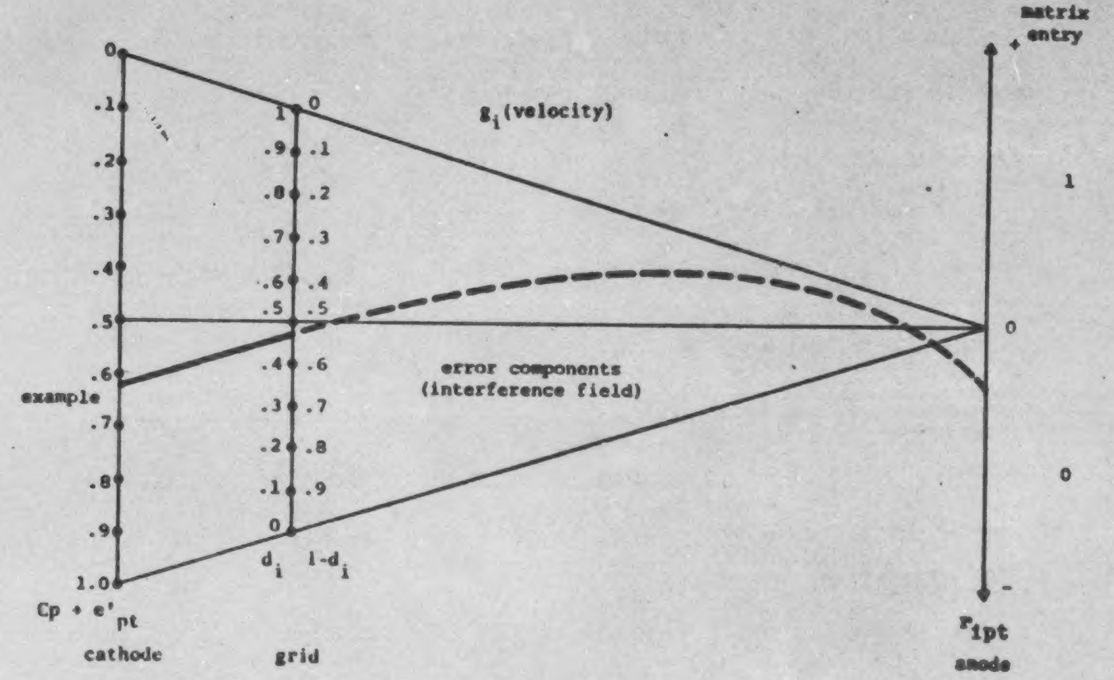

matrix entry
+
0
.1
.2
.3
.4
.5
.6
example
.7
.8
.9
1.0
Cp + e'pt
cathode
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
d_i
1-d_i
grid
g_i (velocity)
error components
(interference field)
1
0
0
-
r_ipt
anode

Figure 3. A cathode-ray tube analogy for the computer program.

termed "initial direction." The particle is emitted with initial velocity g_i and passes through an electromagnetic field of strength $e''_{ip} + e_{ipt}$ toward the anode, r_{ipt}. As Equation 11 indicates, the greater the velocity g_i, the less the effect of the interference field $e''_{pt} + e_{ipt}$.

In the example shown in Figure 3, $c_p + e'_{pt} = .62$ and $1 - d_i = .54$, resulting in an initial direction of $(c_p + e'_{pt}) - (1 - d_i) = .08$. If the velocity of the particle were great enough in comparison to the strength of the interference field, the particle would continue on to the upper, greater-than-zero half of the anode and the entry in the response matrix would be 1. In this example, however, the error components are large enough with respect to g_i to bend the path of the particle downward to the less-than-zero half of the anode, and the entry in the matrix is 0.

It should be further noted that since the computer program is designed to simulate the results of the test-taking process rather than the process itself, the relationships in the computer model among such things as item facility, examinee ability, and test mean are not necessarily those one might expect. For example, test mean is not an algebraic function of item facility alone (as it is in the usual test models), but rather is only influenced by it, and then only in conjunction with person competence (combined with it to produce "initial direction") and subject to the effects of both item goodness and the error components.

The differences between the usual model and the computer model used in this research are due to practical rather than theoretical considerations: the usual model does not readily lend itself to computer simulation since its inner relationships are necessarily bound up with the unpredictability of human behavior. The computer model was evolved over a period of time as the best procedure that the author and his computer-programmer colleague could devise in order to simulate the results of the usual test-taking process.

In Figure 4, the usual relationships (A) and those of the computer model (B) are compared. Arrows indicate directions of relationships, solid lines indicate direct relationships, and dotted lines indicate indirect relationships.

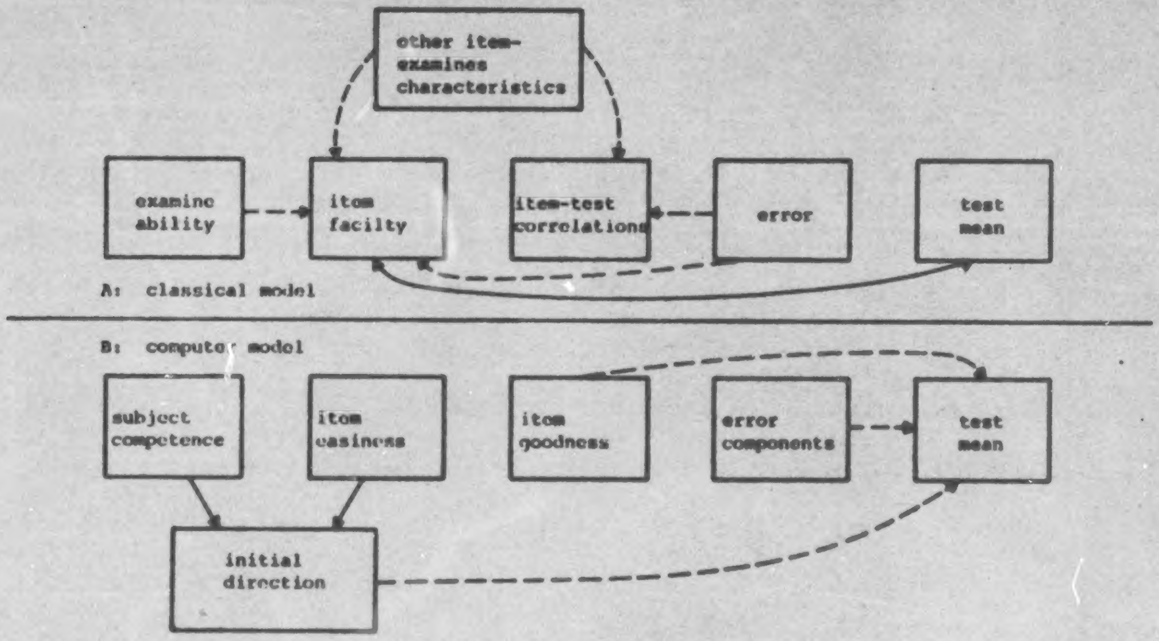

other item-examines characteristics
examine ability
item facilty
item-test correlations
error
test mean
A: classical model
B: computer model
subject competence
item easiness
item goodness
error components
test mean
initial direction

Figure 4. Relationships between item, examinee, and test characteristics; a comparison of the classical (A) and computer (B) models.

Input Parameters

The computer program offers a wide range of options for defining the three major vectors ($\overrightarrow{c_p}$, $\overrightarrow{d_i}$, and $\overrightarrow{g_i}$) (see Appendix C for more detail). However, for the purposes of this study, only a limited variety of options was used.

The competence vector was restricted to two types. One is a normal distribution (Figure 5), with $\mu = .5$ and σ^2 such that all c_p values generated lie between 0 and 1 (explained more fully in Appendix C). This competence distribution was chosen to reflect the classical assumptions about ability within a population.

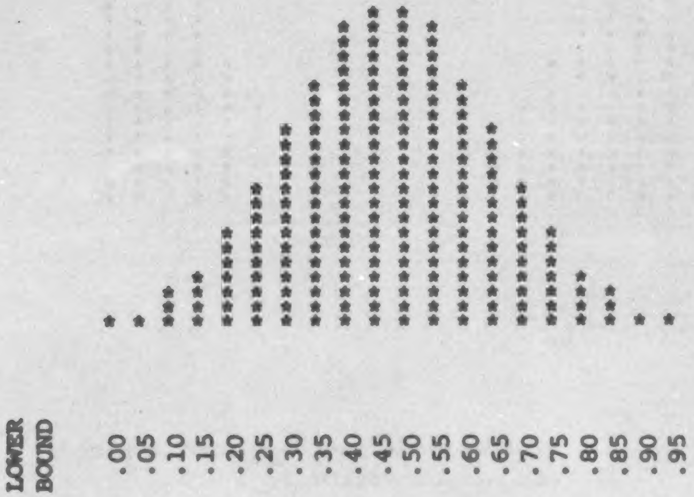


Figure 5. Histogram of components of a normally distributed competence vector ($\overrightarrow{c_p}$)

The second type is a bimodal, "inverse normal" distribution (Figure 6), which is essentially what would be obtained if a normal distribution were cut in half at the center, the left half translated .5 to the right, and the right half translated .5 to the left. This competence distribution was chosen to reflect the notion that, for a given behavioral objective, a student generally either has or has not mastered the objective.

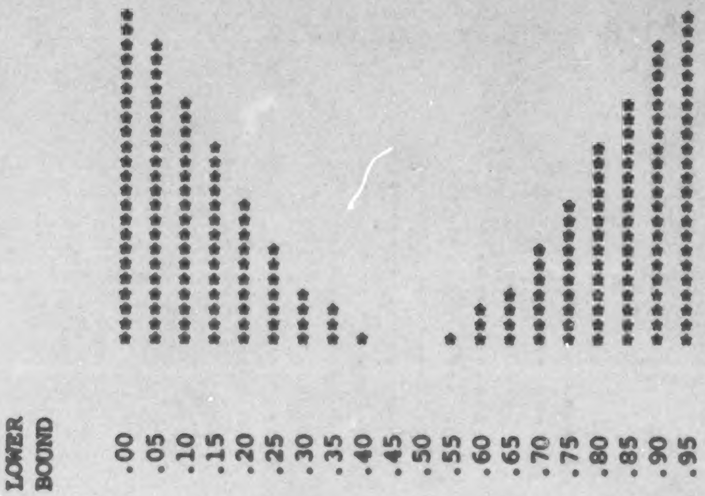


Figure 6. Histogram of components of a bimodal competence vector $(\overline{c}_p)$

The item facility and goodness vectors employed in the study were all uniformly distributed, but their upper and lower bounds varied according to preset conditions.

Eight sets of parameters were used, resulting in eight families of response matrices, score distributions, and test indices. Particular combinations of parameters were chosen to simulate responses to three types of tests.

The first type of test has a moderate number of items, relatively low item goodness, and a wide range of item facilities. It is perhaps best exemplified by a poorly-written teacher-constructed test. Parameter sets 1 and 2, which use the normal and bimodal competence vectors, respectively, are of this type. Examples of the resulting distributions, with some of the accompanying basic test statistics, are given in Figures 7 and 8. These basic test statistics are $\bar{p}$, the test mean expressed as average item difficulty; %V, the variance expressed as a percent of maximum possible variance for an n-item test; S, the index of separation (Equation 6); and r, a classical internal consistency reliability estimate.

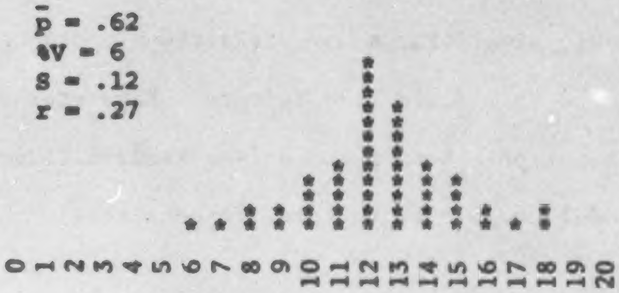


Figure 7: Score distribution resulting from parameter set 1.

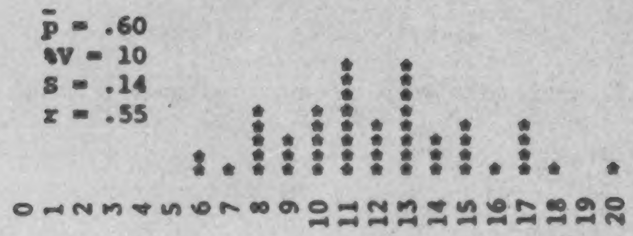


Figure 8: Score distribution resulting from parameter set 2.

The second type of test is short, with relatively high item goodness and a minimal range of item difficulty. It is perhaps best exemplified by a well-constructed criterion-referenced test for a narrow, specific behavioral objective, such as would be found in mathematics. Parameter sets 3 and 4, which use the normal and bimodal competence vectors, respectively, are of this type. (see Figures 9 and 10)

The third type of test is long, with intermediate ranges of item facility and goodness, simulating a more traditional, standardized test, such as would be found in a field like science. Parameter sets 5, 6, 7, and 8 are all of this type. Sets 5 and 6 (see Figures 11 and 12) utilize the normal and bimodal competence vectors, respectively.

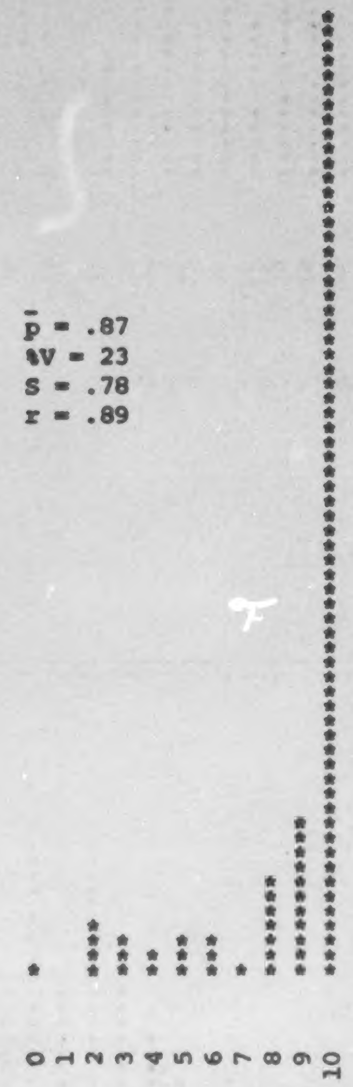


Figure 9: Score distribution resulting from parameter set 3

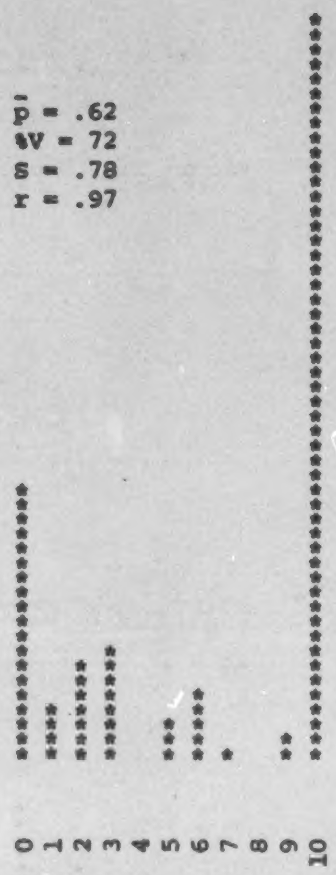


Figure 10: Score distribution resulting from parameter set 4

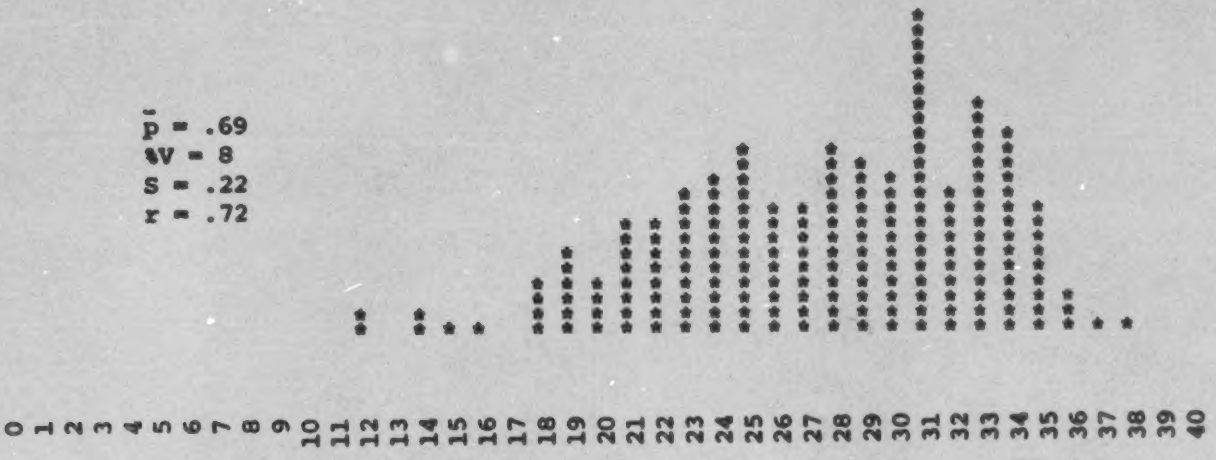


Figure 11: Score distribution resulting from parameter set 5.

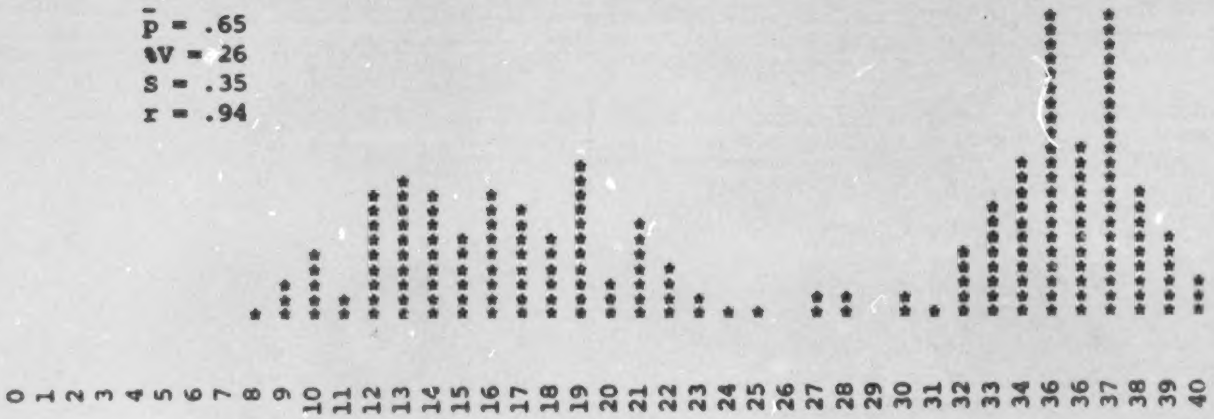


Figure 12: Score distribution resulting from parameter set 6.

For set 7, all parameters are the same as for set 5 except for item facility: the test is more difficult, and hence has generally lower scores and a lower test mean (fee Figure 13).

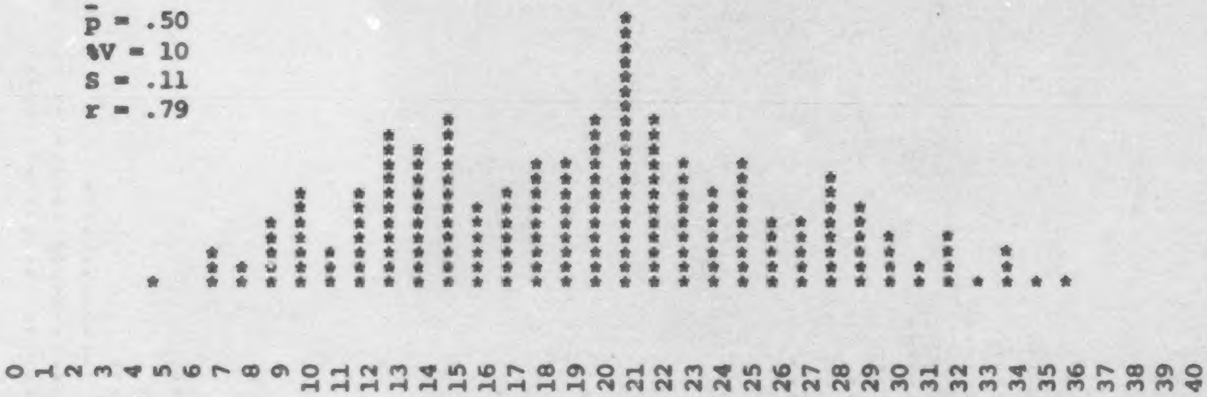


Figure 13: Score distribution resulting from parameter set 7.

Parameter set 8 is the same as set 6 except that the standard deviations of the error components are smaller. This set was chosen because the resulting score distribution (see Figure 14) closely approximates that of empirical score distributions from tests being developed at the Wisconsin Research and Development Center for Cognitive Learning, where this study was conducted.

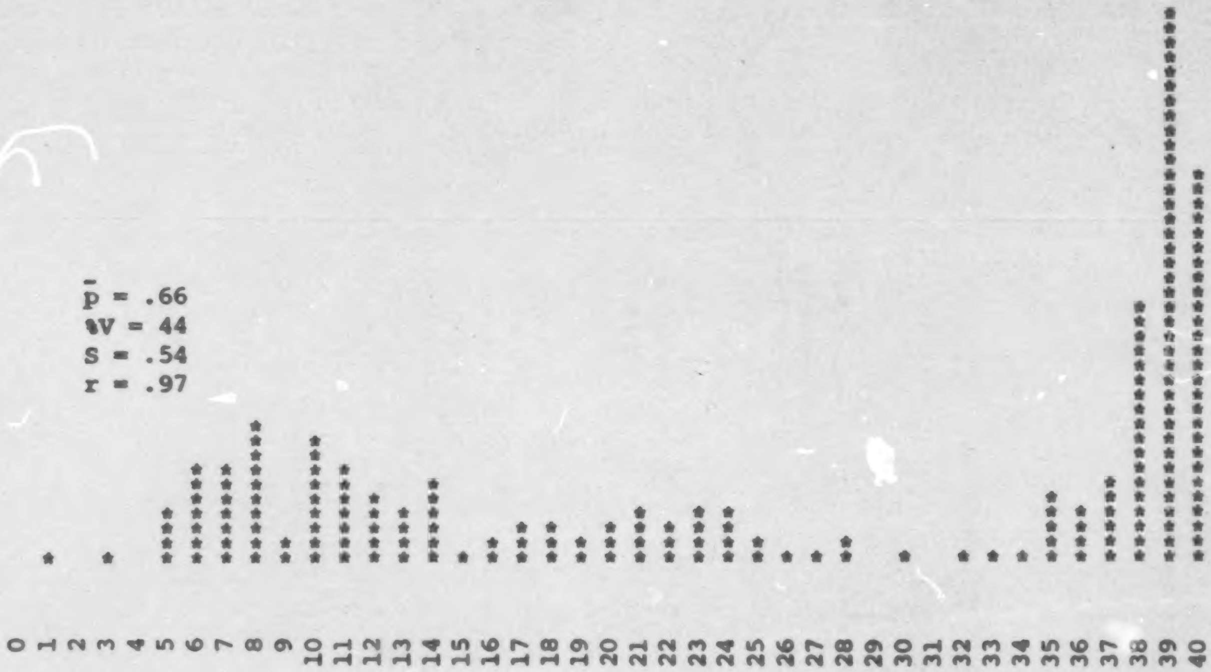


Figure 14: Score distribution resulting from parameter set 8.

Table 4 gives the numerical values used for these eight parameter sets. In all cases, c_p is either normal or bimodal (inverse normal) as described previously, with $\mu = .5$, $c_1 = 0$, and $c_N = 1$; d_i and g_i are uniformly distributed within the intervals shown in the table.

TABLE 4

INPUT PARAMETERS USED FOR THE STUDY

test type	parameter set	$\overrightarrow{c_p}$	$\overrightarrow{d_i}$		$\overrightarrow{g_i}$		error s.d.'s			no. of examinees	no. of items
			min	max	min	max	e'_{pt}	e''_{ip}	e_{ipt}		
1	1	N	.5	1.0	.1	.3	0	.04	.02	50	20
	2	B									
2	3	N	.7	.8	.5	.6	.01	.04	.02	100	10
	4	B									
3	5	N	.6	.9	.2	.4	.01	.04	.02	200	40
	6	B									
	7	N	.35	.65	.2	.4	.01	.04	.02	200	40
	8	B	.6	.9	.2	.4	.01	.02	.01	200	40

The Questions and Research Methods

1. What are the characteristics of coefficient beta? This can be divided into the following questions:

a. What are the values of $\phi(X)$, as defined in Chapter III, for each score X, and for varying numbers of items and criterion levels? Since the algebraic definition of $\phi(X)$ is rather complex, to answer this question graphical displays were made so that the contributions to coefficient beta of $\phi(X)$ for each X could be visually compared. Figure 1 served as an example. The graphs are given in Appendix B.

b. For each parameter set, how does coefficient beta vary as the criterion level changes? To answer this question, response matrices were generated for each of the eight parameter sets, and graphs were drawn. These graphs, and answers to all questions which follow, are given in Chapter VI.

c. What is the behavior of coefficient beta as the number of examinees increases? To answer this, four matrices were generated for each parameter set, using 25, 49, 100, and 400 examinees.

d. What is the behavior of coefficient beta as the number of items increases? Is the Spearman-Brown prophecy formula applicable? To answer these questions, four matrices were generated for each parameter set, using 10, 20, 40, and 80 items, graphs were drawn, and various regression analyses were carried out.

e. Are there predictable relationships between coefficient beta and the following basic test statistics: (1) test mean, expressed as a percent (i.e., mean item difficulty), (2) score variance expressed as a percent of $n^2/4$, the maximum possible variance for a test of n items, (3) index of separation, (4) coefficient alpha (KR-20), (5) KR-21, and (6) percent mastery?

To answer these questions, various analyses, including stepwise analysis of regression, often non-linear, were carried out on the data generated in answer to Question 1b.

2. What are the characteristics of three other criterion-dependent single-administration indices? Harris's index of efficiency, Livingston's criterion-referenced reliability coefficient, and the criterion-referenced index of separation, all discussed in Chapter IV, were computed for the same parameter sets as those for which coefficient beta had been calculated. The analyses were similar to those mentioned under Question 1.

3. Are there predictable relationships between coefficient beta and any or all of these three indices? This question was answered through graphs and analyses of regression.

4. Are there predictable relationships between coefficient beta and other fourfold table indices? The cosine-pi estimate and the phi coefficient (and hence coefficient kappa with equal off-diagonal cells) were calculated for the parameter sets from the table resulting from all possible split-halves. Data were analyzed through graphs and regression analyses.

The Regression Analyses

The regression analysis routine chosen for the study, STEPREG1 (1973), is part of the University of Wisconsin computer center's standard statistical analysis package. The basic purposes of this stepwise analysis of regression program are to analyze the manner (and degree) to which the variance of the dependent variable is explained by variation in the independent variables, and to calculate regression equations. The stepwise feature of this statistical technique allows one to introduce independent variables into the regression equation in any number and in any order, either singly or in groups. If some or all of the variables are allowed to enter as a group, the program determines the magnitude of the contributions of each of these variables toward explaining the variance of the dependent variable and allows these variables to enter the regression equation in order of the magnitude of their contributions. Thus one can analyze not only which independent variables help explain the behavior of the dependent variable, but also which ones are most important. The result can be interpreted as representing a quantified "sociogram" of the indices in the analysis.

Stepwise analyses of regression were used rather extensively in this study because the procedure made it possible to analyze the manner in which coefficient beta and other indices are related to various test statistics and to each other.

CHAPTER VI

RESULTS AND CONCLUSIONS

This chapter is in several sections, roughly corresponding to the questions set forth in the previous chapter. The first section deals with the characteristics of coefficient beta, which was developed in Chapter III, and its relationships to various test parameters and basic test statistics (including classical reliability). The following three sections deal similarly with the three other recently-suggested test indices that were defined and briefly discussed in Chapter IV. The last section discusses the relationships of these four indices among themselves and to the cosine-pi estimate and the phi coefficient defined in Chapter IV.

Characteristics of Coefficient Beta

Values of $\phi(X)$

As mentioned earlier, one approach to the analysis of coefficient beta is to investigate its component parts. Recall from Equation 2 that

$$\beta = \frac{1}{N} \sum_{p=1}^{N} \phi_i(X_p),$$

where N is the number of examinees, X_p is the pth person's total score, and ϕ_i is as defined in Chapter III. Since X_p is a member of the set $\{0,1,...,n\}$, it is useful to inspect the values of $\phi(X)$ for each X in $\{0,1,...,n\}$. Table 5 shows these values of $\phi(X)$ to two decimal places for a 20-item test with a criterion level of 0.7.

TABLE 5

Values of $\phi(X)$ for $n = 20$, $c = .7$

X	0 - 6	7	8	9	10	11	12	13	14	15	16	17 - 20
$\phi(X)$	1.00	.99$^+$	.98	.93	.82	.63	.35	.00	.37	.70	.91	1.00

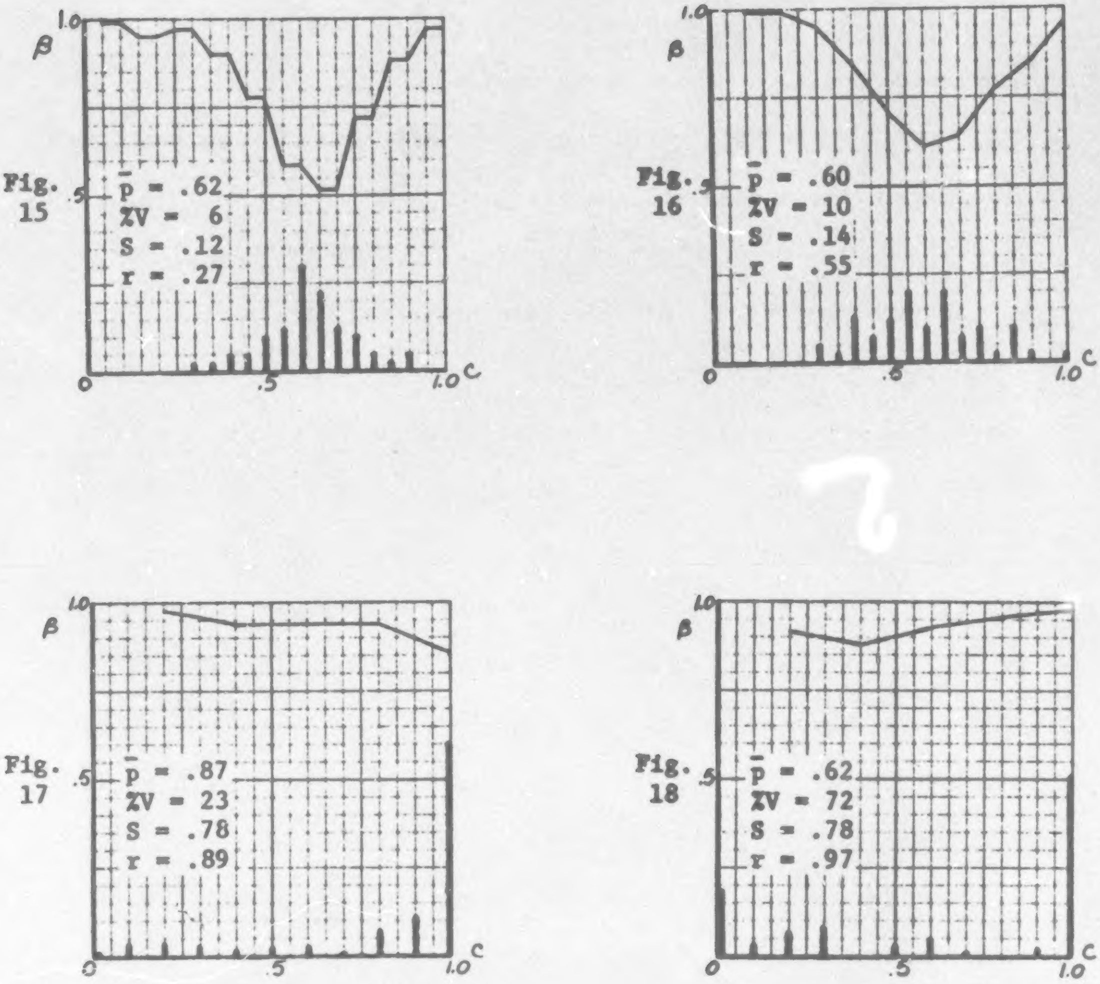
As can be seen, $\phi(X)$ decreases as a person's score nears 13, which is the integer 2k-1 as defined in Chapter III. In general, the farther a person's score is from the cutoff, the greater is $\phi(X)$.

As noted earlier, Figure 1 gives a graph of these values of X and $\phi(X)$. Other graphs of $\phi(X)$, for selected numbers of items and criterion levels, can be found in Appendix B (Figures B1 through B7).

Coefficient beta and criterion level

As described in Chapter ~~V~~, the eight different sets of input parameters selected for the computer program generated eight families of simulated test score distributions. Since the criterion level is an integral part of the formula for coefficient beta, the value of the coefficient will tend to vary as the criterion level changes. Recall that the formula for coefficient beta (see Equation 3) contains k, the minimum score required to achieve a mastery classification on a half-test. Since k of necessity lies in the set $\{1,2,\ldots,n/2\}$ for an n-item test, there are n/2 possible criterion levels and hence n/2 meaningful cutoff scores. Thus, as far as the computation of β is concerned, there are only half as many meaningful criterion levels, and hence half as many values of β, as there are items on the test.

In an actual test situation, the mastery criterion level is unlikely to be less than 0.5, and is perhaps most likely to be in the range [.6,.9]. Nonetheless, for the sake of thoroughness, the values of coefficient beta for all possible criterion levels from 2/n to 1 are shown in Figures 15 through 22 for parameter sets 1 through 8. On each graph, the abscissa is the criterion level and the ordinate is the value of β. For reasons to be discussed shortly, a bar graph of the relative frequency distribution of total scores (see Figures 7 through 14), on the same scale as the criterion level, is also given along the abscissa of each graph. Also included with each graph are certain basic test statistics (defined in the last chapter): $\overline{p}$, the test mean; %V, the percent variance; S, the index of separation; and r, a classical reliability estimate. For the graphs of β, as well as of μ_c^2 and S_c (to be given later), the classical reliability estimate is KR-21, since this statistic is computed from the frequency distribution of total scores as are the three CRT indices. For the graphs of k_{TX}^2 (to be given later), the classical reliability estimate is KR-20, or coefficient alpha, since these statistics are computed from the item-by-person response matrix.



Figures 15-18: Graphs of coefficient beta against criterion level, with score distribution relative frequencies, for parameter sets 1-4.

Fig. 19

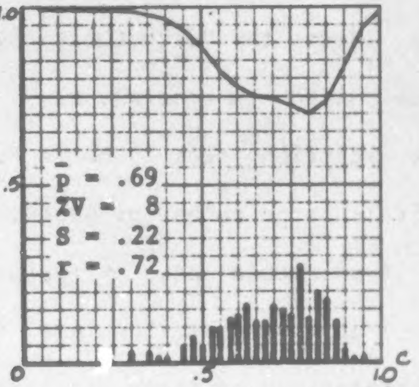


Fig. 20

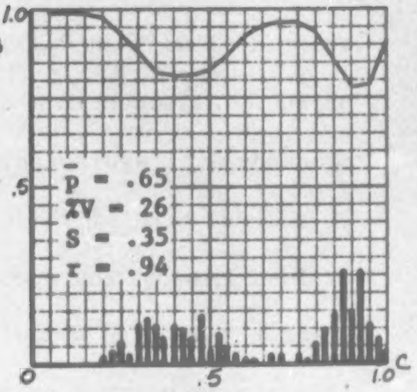


Fig. 21

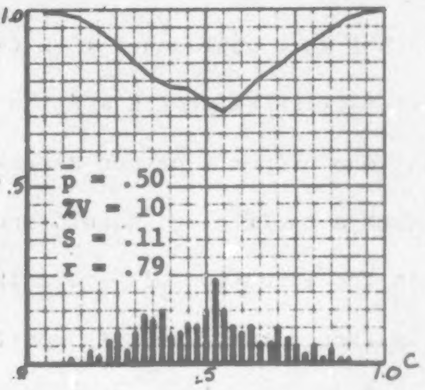
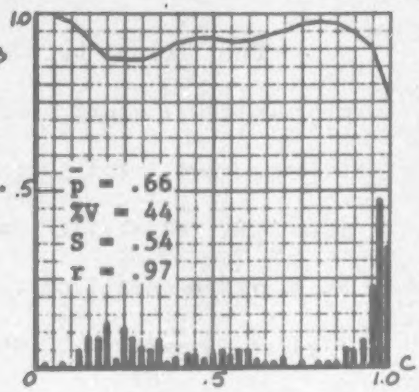


Fig. 22



Figures 19-22: Graphs of coefficient beta against criterion level, with score distribution relative frequencies, for parameter sets 5-8.

The graphs show that as c approaches 0, β approaches 1. This limiting value is reasonable since a criterion level of 0 "separates" those examinees with a score of 0 or more from those examinees with a score of less than 0, an impossibility. Hence the "dichotomization" is perfect, although in a degenerate sense. Also, in general, as c approaches 1, β again approaches 1. The exceptions seem to be in Figures 17 and 22, both of which show a relatively large number of scores slightly less than n, the number of items. If one could set a criterion level greater than 1, β would take the value 1 at that criterion level since, like the case $c = 0$, $c > 1$ implies "separation" of those examinees with scores greater than n (another impossibility) from those examinees with scores less than or equal to n.

Coefficient beta and the score distribution

Coefficient beta does not approach 1 as c approaches 1 in the graphs of Figures 17 and 22 because of the interaction between β and the distribution of total scores. Recall that one property deemed desirable for a CRT reliability index was that such a coefficient should increase as scores depart from the cutoff. With the exception of Figure 18, Figures 15 through 22 show that this is indeed the case with coefficient beta, although these graphs show this relationship in another way; in these figures, the coefficient increases not as the scores depart from the cutoff, but as the cutoff departs from the mode(s) of the score distribution. Perhaps the best examples of this phenomenon are shown in Figures 19 through 22, where there are more items on the test and thus smoother curves of β values.

Note, however, that the curve of β values "lags behind" the bar graph representing the frequency distribution of scores. This lag is most easily discerned on the graphs with clearly defined score distribution modes: Figures 15, 19, 21, and 22. In these instances the criterion level corresponding to a cutoff score immediately above the mode(s) yields the minimum value(s) of the coefficient. The lag is due to the fact that the score that contributes zero to β is 2k-1, one less than the cutoff. This explains why β does not have its minimum value at the mode in Figure 18, and why it does not drop as sharply as one might expect at the mode in Figure 17.

At any rate, it is clear that the shape and modes of the score distribution in relation to the cutoff have an important effect on the value of β.

Coefficient beta and basic test statistics

The basic test statistics considered in this section are those given in Figures 15 through 22 and described earlier. They are invariant for a given item-by-person response matrix; they do not change as the criterion level varies. The data available for this and later statistical analyses include values of β at all possible criterion levels for 24 score distributions: 3 representatives of each of the eight distribution types. Since β, unlike the basic test statistics, varies as criterion level varies, it is not meaningful to include all data points in an analysis comparing β to these basic test statistics. One can, however, investigate the relationship if the variance in β due to the changing criterion level is removed. This can be done in one of (at least) two ways: by taking either the minimum or the mean value of β over all criterion levels for a given score distribution. Table 6 shows the rank order of the eight distributions on each basic test statistic, as well as on min(β) and $\bar{\beta}$.

TABLE 6

ORDINAL RANK OF EACH DISTRIBUTION ON THE VARIABLE
INDICATED AT TOP OF COLUMN
1 = LOW, 8 = HIGH

Distribution from Fig. No.	$\bar{p}$	%V	S	KR-21	min β	$\bar{\beta}$
15	3	1	2	1	1	1
16	2	3	3	2	2	2
17	8	5	7	5	7	6
18	4	8	8	7	8	8
19	7	2	4	3	3	4
20	5	6	5	6	6	5
21	1	4	1	4	4	3
22	6	7	6	8	5	7

From the data in Table 6, Spearman's rank-order correlation (ρ) was computed for both min(β) and $\bar{\beta}$ against each of the four basic test statistics. Table 7 presents these computed values of ρ. The computed ρ is at least as high for $\bar{\beta}$ as for min(β) in each case.

TABLE 7

VALUES OF SPEARMAN'S RHO (RANK-ORDER CORRELATION) BETWEEN MIN β, $\bar{\beta}$, AND BASIC TEST STATISTICS

	$\bar{\rho}$	%V	S	KR-21
min β	.43	.88	.83	.83
$\bar{\beta}$	.55	.90	.90	.93

Test mean appears to have little to do with coefficient beta. The best correspondence seems to be that of KR-21 with $\bar{\beta}$. However, it should be pointed out that other test indices, which are analyzed later, correspond about equally well with some of the same basic test statistics.

Coefficient beta and the number of examinees

For a given set of test parameters and a given criterion level, variation in the number of examinees does not seem to have any systematic effect on the value of β. Figure 23 is a scatterplot of β for 2N (or, in some cases, 4N) examinees against β_N. The pairs of numbers used[3] were (25,49), (49,100), and (100,400). The correlation of β_N and β_{2N} (or β_N and β_{4N}) was high, .94. The obtained linear regression equation was $\widehat{\beta}_{2N} = -.001154 + .9995\beta_N$, which is very close to the model $\beta_{2N} = \beta_N$. In fact, the fit is close enough to allow one to assume without qualm that the model obtains in the population. This result was expected, since $\beta = \frac{1}{N} \sum_X f_x \phi_i(X)$, and hence doubling the number of examinees should merely tend to double each f_x (as well as double N), resulting in algebraic cancellation.

---

[3] The number 49 was chosen in place of the perhaps more obvious 50, on the chance that there was a connection between $\sqrt{N}$ and β. Results showed there was no such connection.

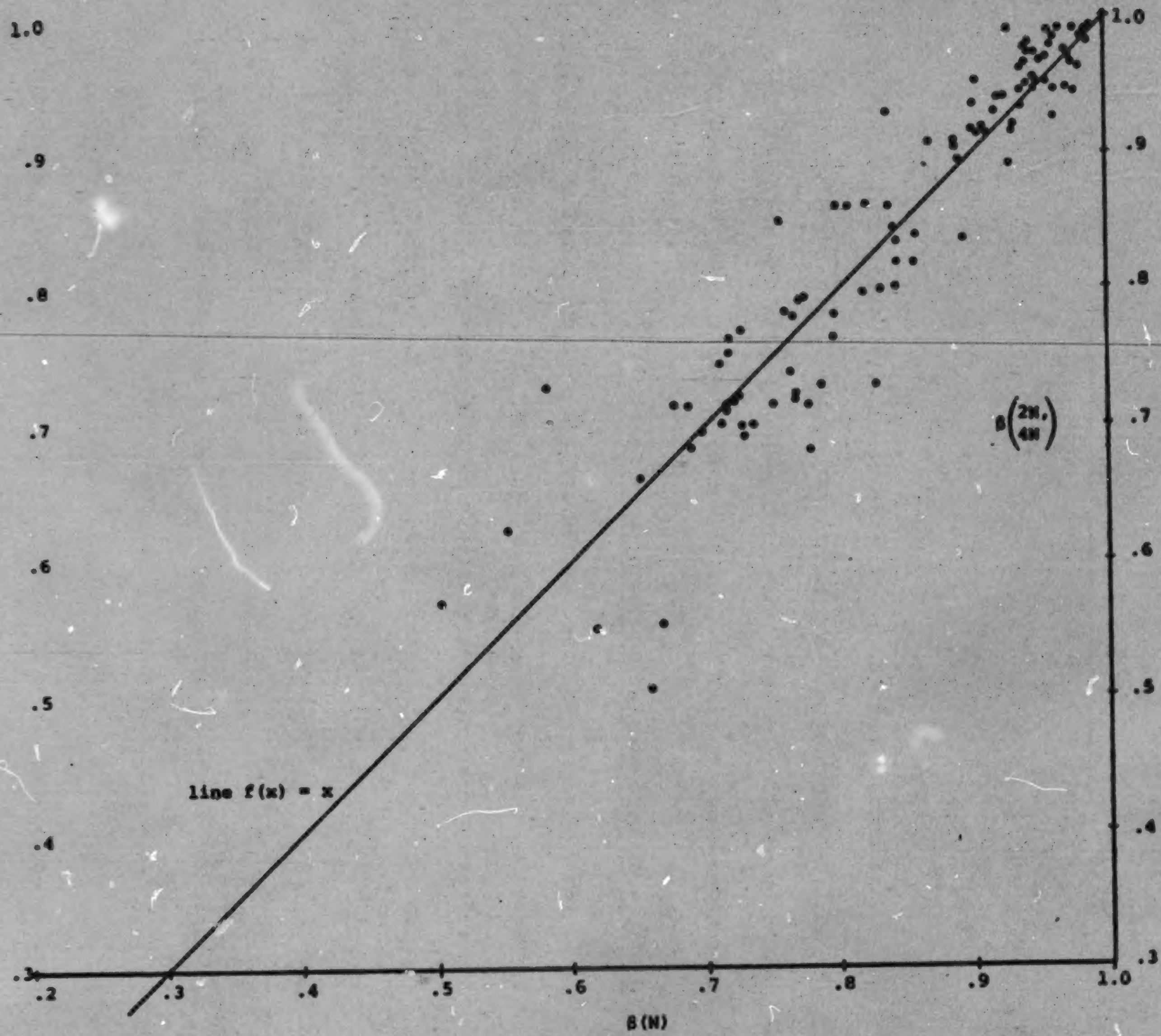


Figure 23. Scatterplot of β for 2N (or 4N) examinees against β for N examinees.

Coefficient Beta and the Number of Items

For a given set of test parameters and a given criterion level, variation in the number of items affects the value of β: in general, β increases as the number of items increases. Figure 24 is a scatterplot of β for 2n items again β for n items. (For comparison purposes, the stars in the figure represent values of coefficient alpha: α_{2n} against α_n). For this figure, (n,2n) took on values of (10,20) (20,40) and (40,80). The scatterplot incorporates all data calculated for all criterion levels on eight score distributions, one from each of the eight parameter sets. The linear correlation of this set of points (considered as a set of ordered pairs) is quite high, .939.

Figure 24 shows two curves. The lower one is the line $\beta_{2n} = \beta_n$, i.e., what would be expected if the number of items had no effect on the value of β (henceforth called the N-E line). The figure shows fairly clearly that most of the points are above the N-E line rather than evenly distributed around it. The regression equation was $\hat{\beta}_{2n} = .1999 + .8151\beta_n$, consistent with the observation that most of the scatterplot points lie above this line. In this case the coefficient of determination (the squared correlation, and thus the percent of variance of β_{2n} accounted for by variance in β_n) is .881. That is, 88% of the variance exhibited in the values of β_{2n} can be explained by the model $\beta_{2n} = \beta_n$.

The upper curve in Figure 24 is $\beta_{2n} = \frac{2\beta_n}{1+\beta_n}$, i.e., the graph that would be expected if the Spearman-Brown prophecy formula held (henceforth called

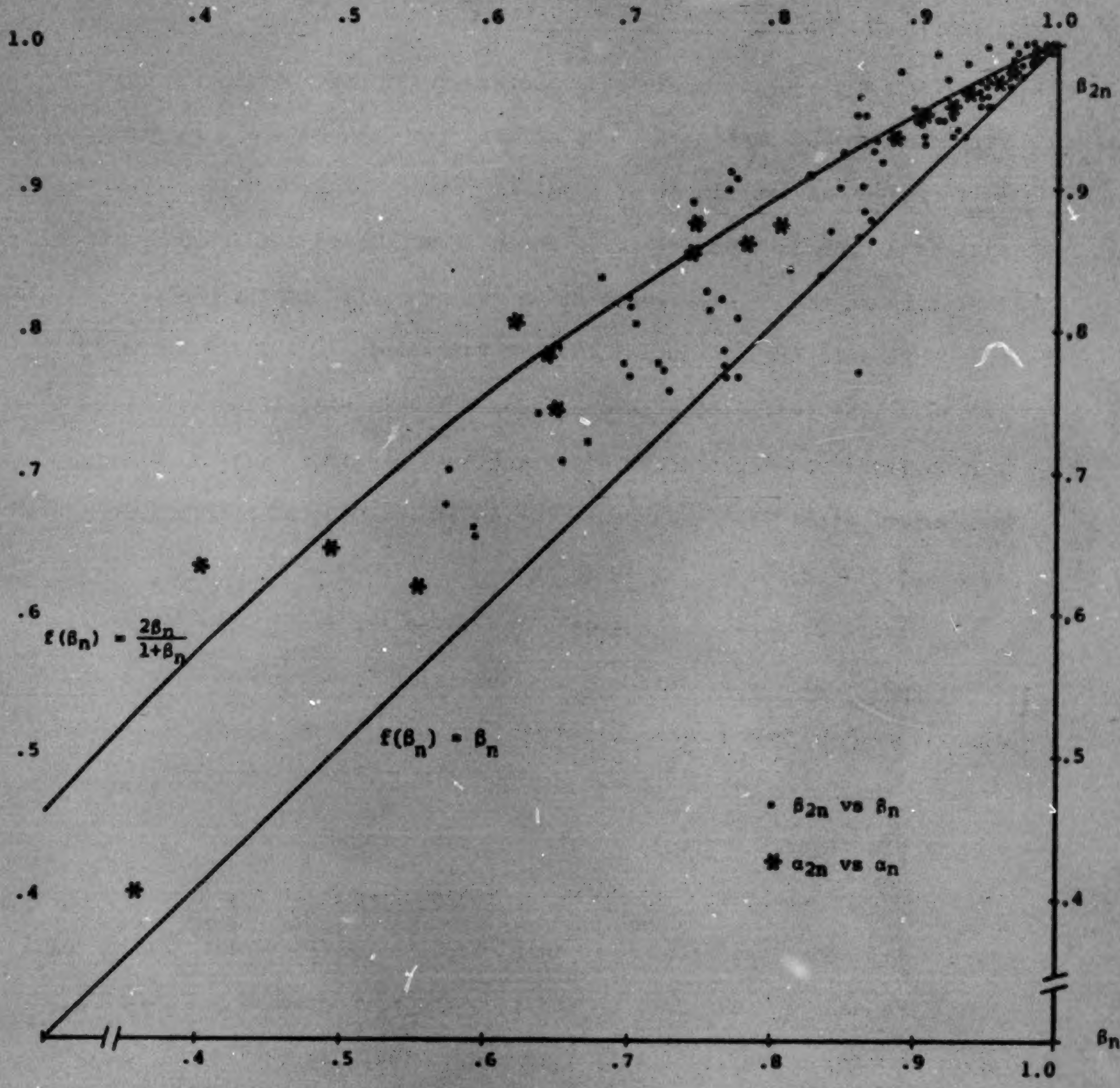


Figure 24. Scatterplot of β (and α) for 2n items against β (and α) for n items.

the S-B curve). At first glance this would appear to be a better model than the lower line. Yet the points are not evenly distributed around the upper curve: more points are below it than above it. The regression equation for this model was $\hat{\beta}_{2n} = -.3015 + 1.307 \frac{2\beta_n}{1+\beta_n}$, consistent with the observation that more points are below the curve than above it. The coefficient of determination for this model was .887, only minimally higher than that for the linear no-effect model. Hence the Spearman-Brown model does not appear to explain the behavior of β better than the no-effect model. Nonetheless, using the evidence presented here, one could claim that the former model does at least as well as the latter.

It is illuminating to put aside the computer-generated data for the moment and briefly investigate the behavior of β for some theoretical score distributions: normal, uniform, and symmetric ("inverse normal") bimodal distributions. If for each distribution β_{2n} is plotted against β_n, there appears to be a pattern. Figures 25, 26, and 27 are scatterplots for the normal, uniform, and bimodal distributions, respectively.

Notice that for a normal distribution (Figure 25), the points (β_n, β_{2n}) are approximately evenly distributed about the Spearman-Brown curve $\beta_{2n} = \frac{2\beta_n}{1+\beta_n}$ and none falls below the no-effect line $\beta_{2n} = \beta_n$.

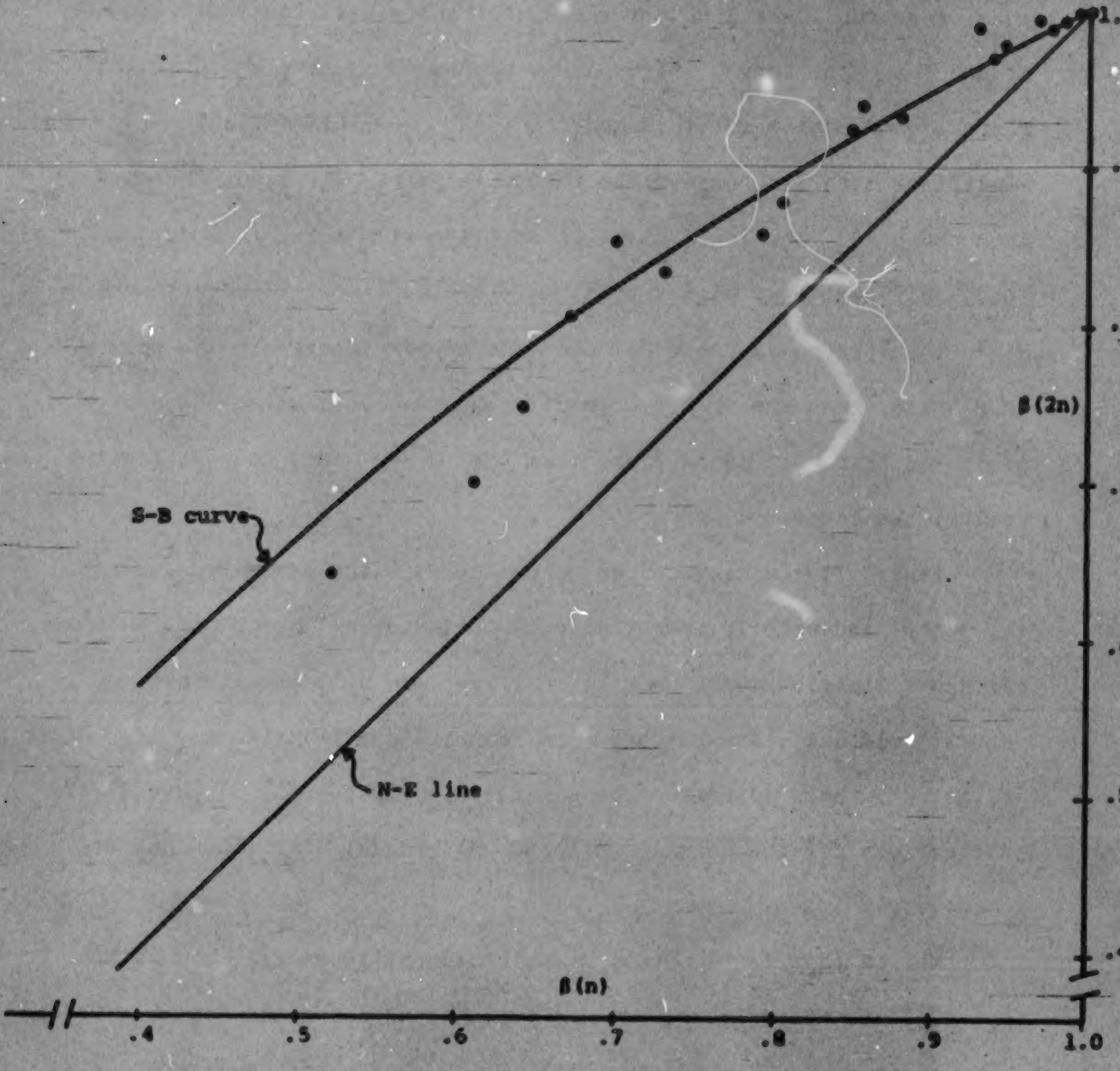


Figure 25. Scatterplot of β for 2n items against β for n items, for a normal distribution.

In fact, none seems to fall below an imagined curve halfway between the S-B curve and the N-E line. Such a half-way curve can be generated by

$$\beta_{2n} = \frac{1}{2}\left[\frac{2\beta_n}{1+\beta_n} + \beta_n\right]$$

$$= \frac{\beta_n(3+\beta_n)}{2(1+\beta_n)} \qquad [12]$$

In the case of the uniform distribution (Figure 26), all points lie between the S-B curve and the half-way curve just described. And, although this figure does not show it, the data from which the figure was drawn indicate that the points lie at or near the half-way curve when the criterion level is near .5 and approach the S-B curve when the criterion level is 1.

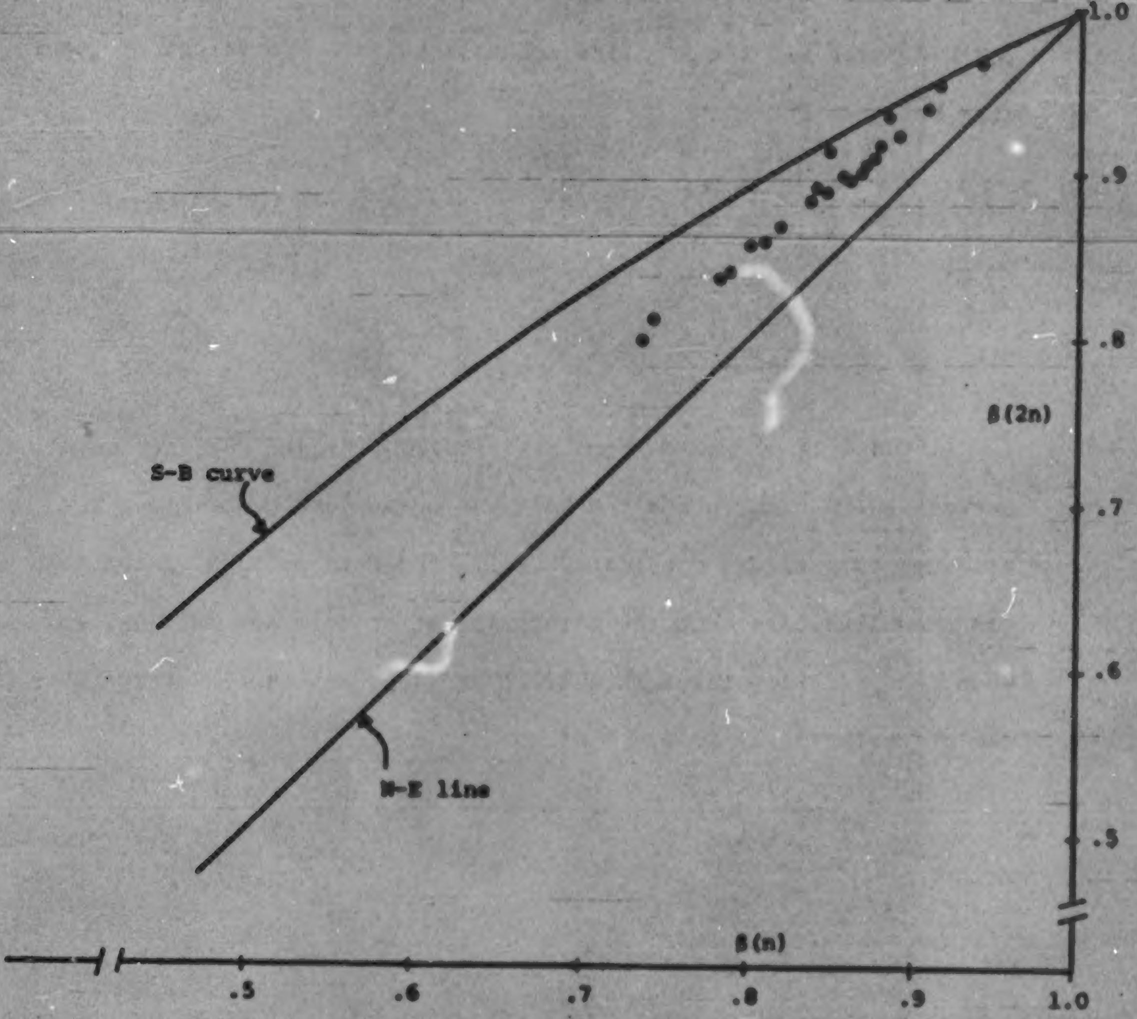


Figure 26. Scatterplot of β for 2n items against β for n items, for a uniform distribution.

For the bimodal distribution (Figure 27), all points appear to lie on or between the S-B curve and the N-E line. Therefore, the value of coefficient beta is apparently affected by the number of items on the test: the more the items, the higher the value of β for a given criterion level and test type.

The shape of the score distribution seems to have some bearing on whether the no-effect model ($\beta_{2n} = \beta_n$) or the Spearman-Brown model ($\beta_{2n} = \frac{2\beta_n}{1+\beta_n}$) holds: for a sharply bimodal distribution, both models seem to account for the variance equally well; for a low-variance normal distribution, the Spearman-Brown model appears to account for the variance better than does the no-effect model.

Interestingly, the computer-generated data follow very closely the half-way curve model described previously. An analysis of regression (of β for 2n items against $\frac{\beta(3+\beta)}{2(1+\beta)}$ for n items) yielded a coefficient of determination of .884, about the same as for the earlier two, and a regression equation of $\hat{\beta}_{(2n)} = .00633 + 1.005\ \beta_{H(n)}$, where $\beta_H = \frac{\beta(3+\beta)}{2(1+\beta)}$. Unlike the earlier two, this regression equation is so near to $\beta(2n) = \beta_H(n)$ that one is tempted to hypothesize that the half-way curve is the appropriate model for the population, and that it should replace both the Spearman-Brown prophecy formula and the no-effect model as far as β is concerned. (It may also be, of course, that any appropriate prophecy formula must come from a totally different framework. This possibility is discussed briefly in Chapter VII.)

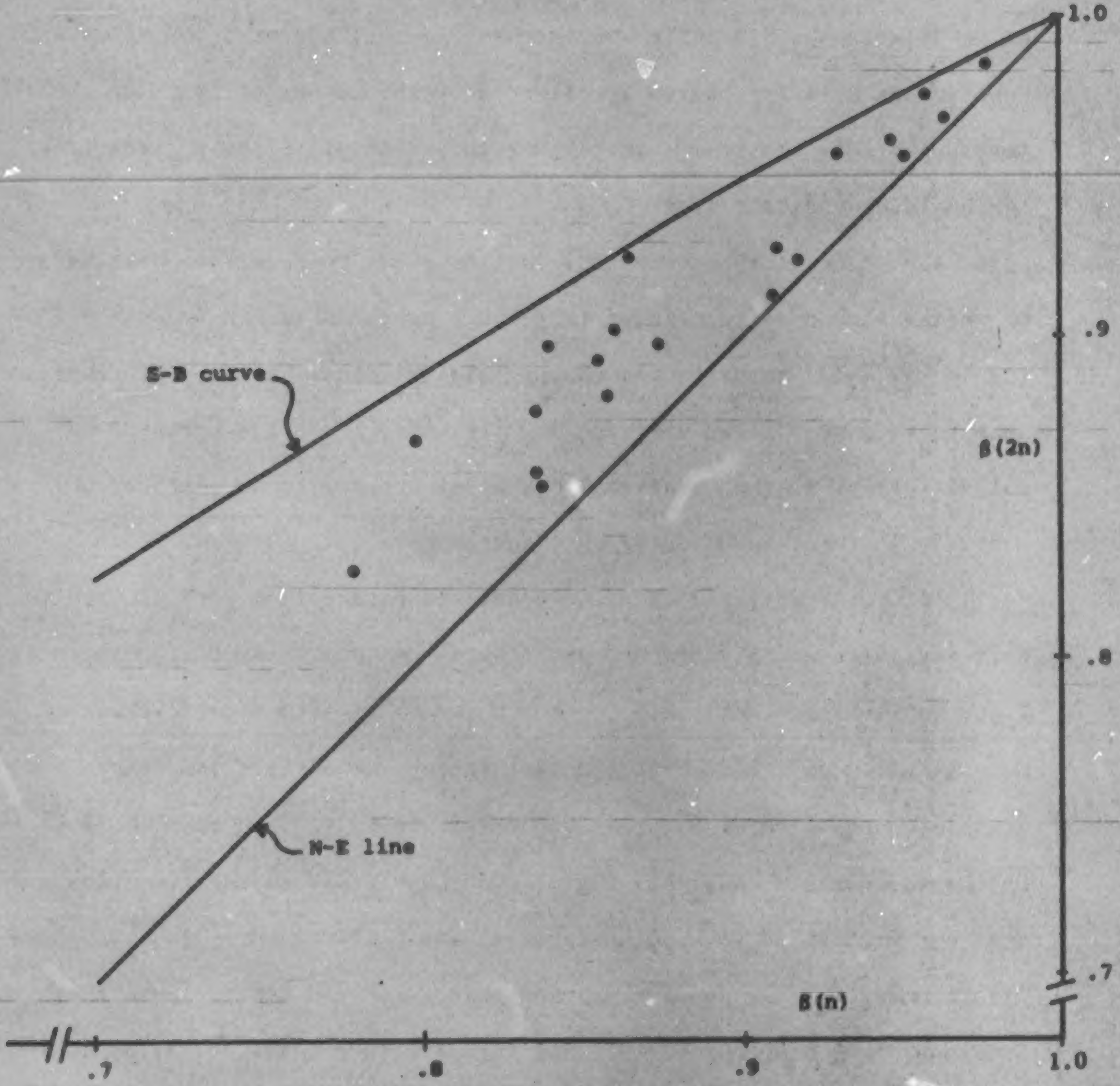


Figure 27. Scatterplot of β for 2n items against β for n items, for a bimodal distribution.

Characteristics of Livingston's k^2_{TX}

Livingston's k^2_{TX}, unlike coefficient beta, is not additive, and thus there is no parallel with the $\phi(X)$ analysis presented for β. There are, however, other parallels between the two indices. As these comparisons are discussed in the last section of this chapter, this section will be concerned only with the characteristics of k^2_{TX}.

k^2_{TX} and criterion level

The computing formula for k^2_{TX}, which shows its relationship to other test statistics, was given earlier (Equation 4) as

$$k^2_{TX} = \frac{r\sigma^2 + (\bar{X}-C)^2}{\sigma^2 + (\bar{X}-C)^2} \quad . \qquad [13]$$

Thus k^2_{TX} has (usually) a different value for each value of C, the cutoff point. In fact, unlike coefficient beta, the number of values of k^2_{TX} for a given item-by-pupil response matrix is limitless since C need not be an integer (Livingston, 1972a). For this investigation, however, values of C were restricted to the set {.05n, .10n; .15n, ..., 1.0n} where n is the number of test items, the same (where meaningful) as for coefficient beta.

The graphs in Figures 28 to 35 show the value of k^2_{TX} at the selected criterion levels for the representatives of the eight score distributions. As before, the relative frequency distribution of total scores, on the same scale as the criterion level, is included with each graph, along with the basic test statistics.

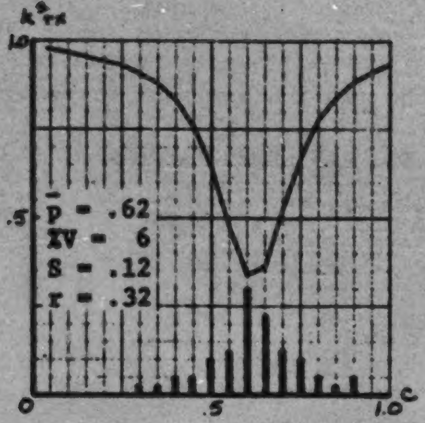


Fig. 28

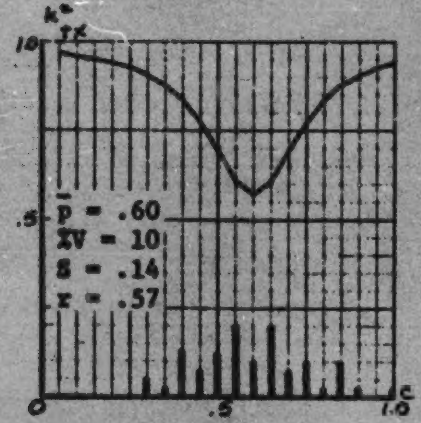


Fig. 29

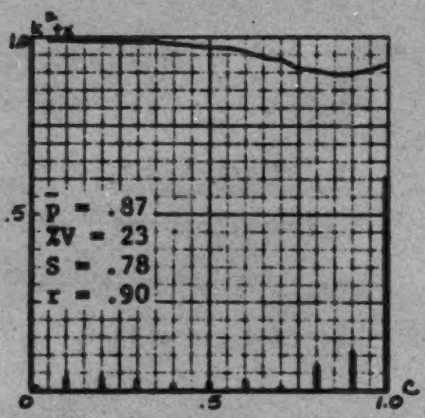


Fig. 30

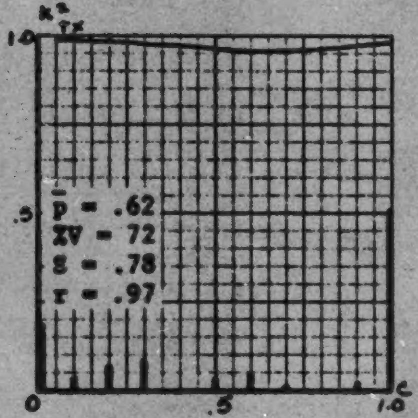


Fig. 31

Figures 28-31: Graphs of k^2_{TX} against criterion level, with score distribution relative frequencies, for parameter sets 1-4.

Fig. 32

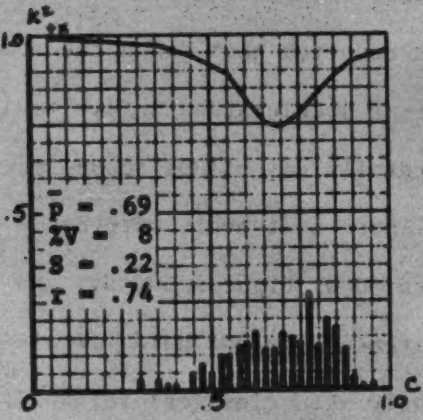


Fig. 33

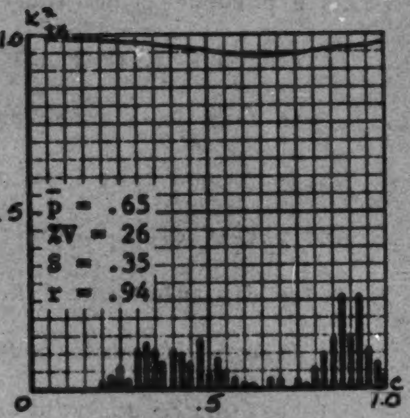


Fig. 34

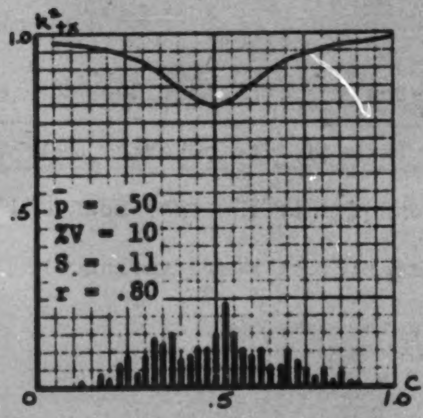
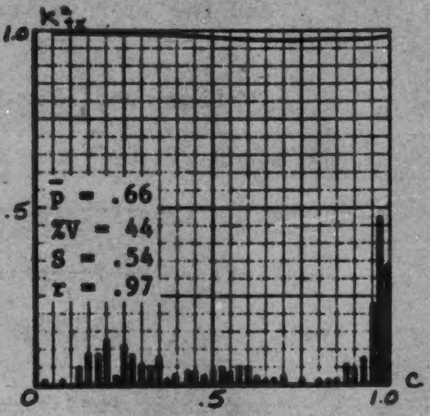


Fig. 35



Figures 32-35: Graphs of k^2_{TX} against criterion level, with score distribution relative frequencies, for parameter sets 5-8.

The graphs indicate that, as would be expected from Equation 13, k^2_{TX} has a minimum at the criterion level nearest the test mean (expressed as a percent), and increases as C departs from the mean. In earlier research on k^2_{TX} (Marshall, 1973) it was reported that "when the mean departs from the criterion, the coefficient accelerates rapidly toward unity," and that "the coefficient generally has values above .95, and rarely drops below .90 [p. 14]." These statements were based on score distributions like those represented by Figures 30, 31, and 35. As figure 28 shows, however, these statements do not hold for all kinds of test score distributions, particularly when classical reliability is low.

k^2_{TX} and the score distribution

Unlike coefficient beta, Livingston's coefficient does not reflect the modes of the score distribution. Instead, its behavior over changing criterion levels seems to be a function of only the test mean and the classical reliability (and thus, indirectly, of score variance). Again, formula 13 indicates that this must be the case.

k^2_{TX} and basic test statistics

Two relationships, both of which follow directly from formula 13, hold true for k^2_{TX}: the minimum value of k^2_{TX} (if the curve were made continuous) is the same as KR-20, and this minimum value always occurs at the test mean. It follows that the rank-order correlation of the minimum value (over criterion levels) of k^2_{TX} with KR-20 is unity.

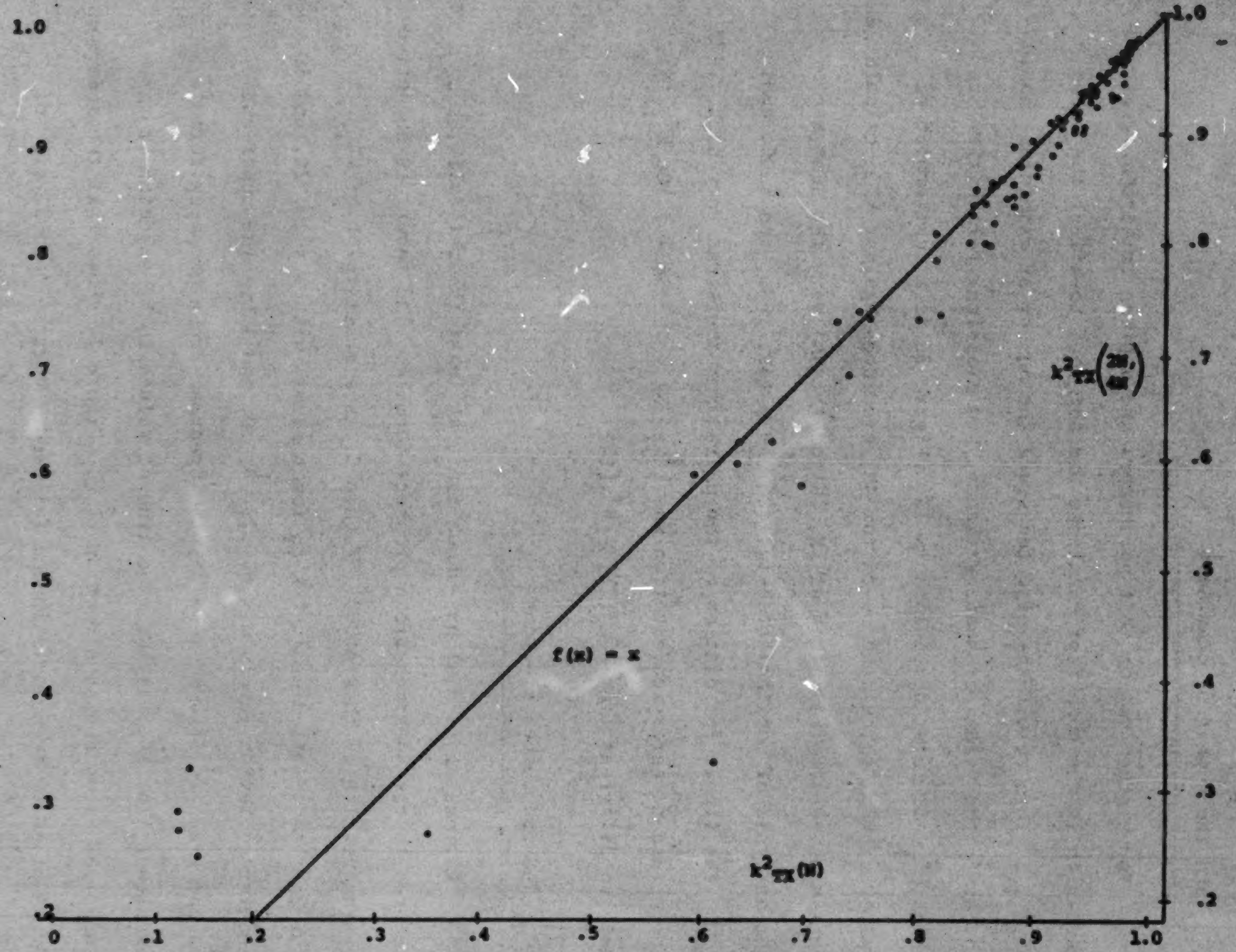


Figure 36. Scatterplot of k^2_{TX} for 2N (or 4N) examinees against k^2_{TX} for N examinees.

k^2_{TX} and the number of examinees

For a given set of test parameters and a given criterion level, variation in the number of examinees did not seem to affect the value of k^2_{TX}. This result was expected since the number of examinees should not alter the values of mean, variance, and classical reliability, to which k^2_{TX} is related via formula 13. Figure 36 is a scatterplot of values of k^2_{TX} calculated on 2N (or 4N) examinees against k^2_{TX} calculated on N examinees. The values of (N, 2N) or (N, 4N) were the same as for the analysis of coefficient beta: (25, 49), (49, 100), and (100, 400).

The linear correlation of the pairs of numbers was very high, .978. The obtained regression equation was $k^2_{TX}(2N) = -.1056 + 1.106\ k^2_{TX}(N)$, not too different from $k^2_{TX}(2N) = k^2_{TX}(N)$.

k^2_{TX} and the number of items

Livingston (1972a) has shown that, at least theoretically, k^2_{TX} adheres to the Spearman-Brown prophecy formula. The theory is supported by the results of this study. Figure 37 is a scatterplot of k^2_{TX} for 2n items plotted against k^2_{TX} for n items, with n = 10, 20, and 40, and for criterion levels of .6, .7, .8, .9, and 1.0. The upper curve on the graph is $f(x) = \frac{2x}{1 + x}$, the Spearman-Brown prophecy formula; the lower line is $f(x) = x$, the line of values to be expected if the number of items has no effect on k^2_{TX}. Figure 37 shows that the Spearman-Brown prophecy formula is indeed followed. Regression analysis (of k^2_{TX} for 2n items against a stepped-up k^2_{TX} for n items) yielded a rather high coefficient of determination of .94 and a regression equation of

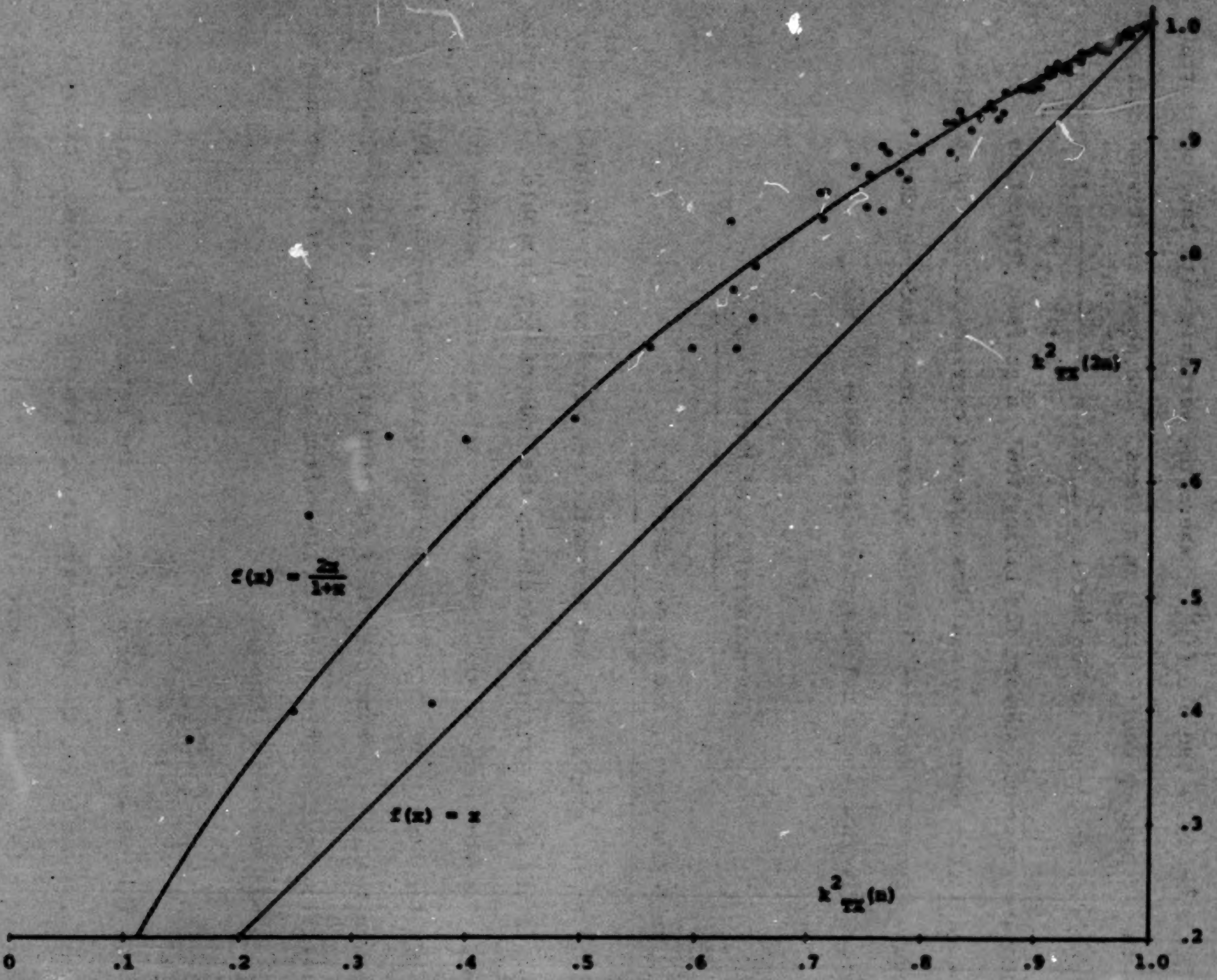


Figure 37. Scatterplot of k^2_{TX} for 2n items against k^2_{TX} for n items.

$\hat{k}^2_{TX}(2n) = .095 + .90\ k^2_{TX}(n)^{S-B}$, where the variable with the superscript is the stepped-up coefficient for n items. The above regression equation is close enough to the model $k^2_{TX}(2n) = \dfrac{2k^2_{TX}(n)}{1 + k^2_{TX}(n)}$ to give moderate empirical support for Livingston's algebraic derivation. Although linear regression analyses were not carried out for the no-effect model, Figure 37 suggests that the Spearman-Brown model produces a much better fit than would a linear no-effect model.

Characteristics of Harris's μ^2_c

μ^2_c, criterion level, and percent mastery

In the graphs for each parameter set given earlier in this chapter for β and k^2_{TX}, criterion level was the independent variable. Criterion level was not used for the independent variable in the graphs for μ^2_c since the results of this study and an earlier one (Marshall, 1973) showed that μ^2_c is more clearly a function of percent mastery than of criterion level. This result follows from an analysis of the formula for μ^2_c, given earlier as Equation 5:

$$\mu^2_c = \frac{SS_b}{SS_b + SS_w} ,$$

where the terms in the ratio represent the between-group and within-group sums of squares for the groups resulting from the dichotomous classification of a CRT. If two or more criterion levels yield the same percent mastery, there is no change in μ^2_c. No matter what the criterion level, if there is only one classification (i.e., if one of the groups has no members), $SS_b = 0$ and hence $\mu_c^2 = 0$, provided there is some score variance within the non-empty group. Hence μ_c^2 always approaches 0 as the percent mastery approaches 0 or 1.

Thus in Figures 38 through 45, percent mastery rather than criterion level is the independent variable. As Harris (1972a) points out, there are as many sortings into groups, and hence values of percent mastery, as there are test scores with a frequency of one or more in the score distribution.

Figures 38 through 45 show that the curve for μ_c^2 as a function of percent mastery is quite smooth and clearly monotonic on either side of the maximum value of μ_c^2. In fact, it appears that one could concoct a non-linear algebraic function of percent mastery (perhaps with some additional variables) that would fit the points precisely. Some attempts were made during this study to construct such a function. Although some functions yielded a close fit, an exact fit was not achieved. These findings will shortly be discussed further.

μ_c^2 and the score distribution

There appeared to be no relationship between μ_c^2 and the score distribution, at least not in the way that the value of β reflects the score distribution mode(s), although the maximum value of μ_c^2 often occurred near the point where there was 50% mastery.

Fig. 38

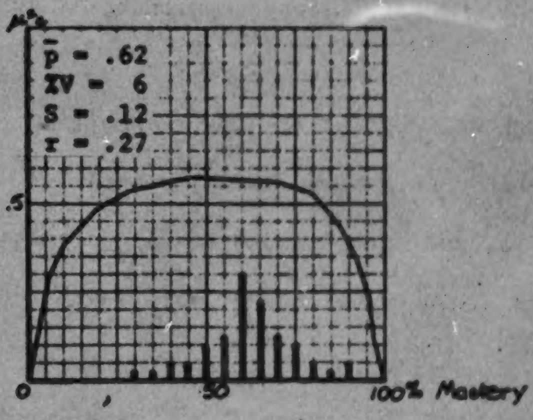


Fig. 39

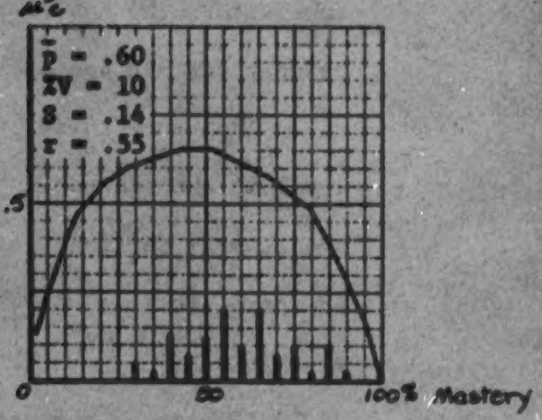


Fig. 40

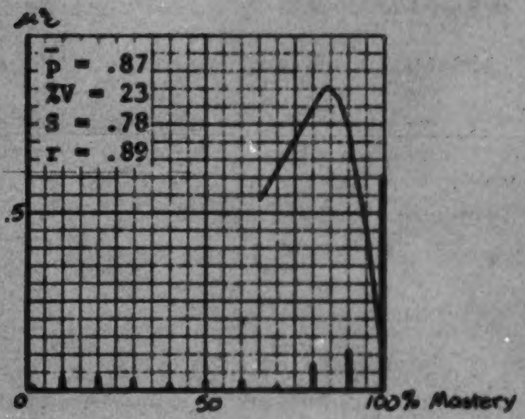
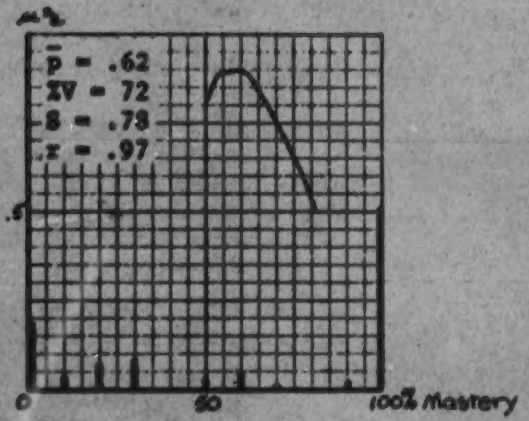


Fig. 41



Figures 38-41: Graphs of μ_c^2 against percent mastery, for parameter sets 1-4.

Fig. 42

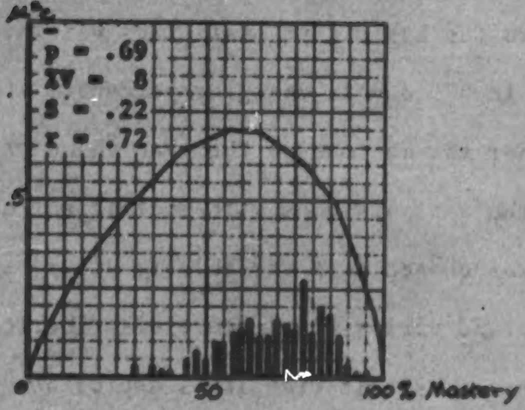


Fig. 43

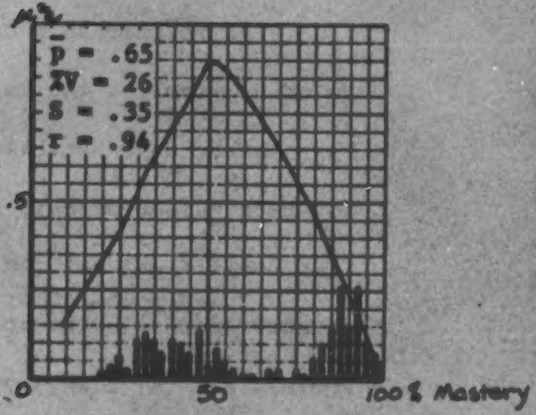


Fig. 44

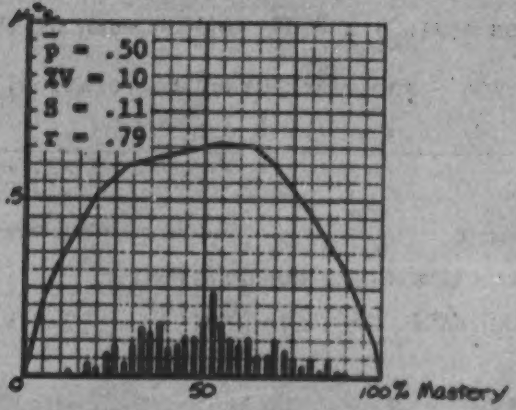
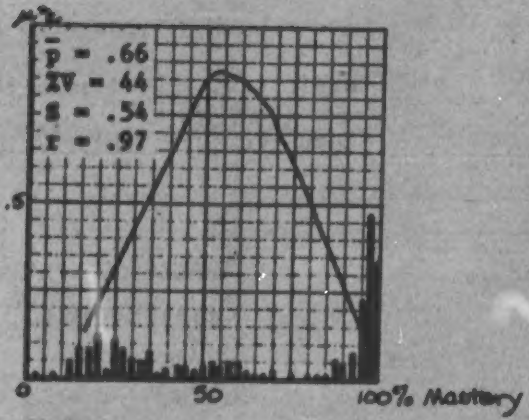


Fig. 45



Figures 42-45: Graphs of μ_c^2 against percent mastery, for parameter sets 5-8.

μ_c^2 and basic test statistics

The relationship of μ_c^2 to the basic test statistics was investigated by removing the variance in μ_c^2 due to changing percent mastery. This can be done by taking either the maximum or the mean value of μ_c^2 for each parameter set, as in the analysis done with coefficient beta. (For coefficient beta, min(β) was chosen as a variable to study because it corresponds to the modes of (and varies over) score distributions, whereas max (β) always approaches 1 as criterion level approaches 0 -- see Figures 15 through 22. For μ_c^2, max(μ_c^2) was chosen instead because it varies over score distributions, whereas, except for the truncated distributions shown in Figures 40 and 41, min(μ_c^2) always approaches or reaches 0 as percent mastery approaches its extremes.) The eight score distribution types were ranked on max(μ_c^2) and $\overline{\mu_c^2}$ and on each of the basic test statistics, and Spearman's rho (rank-order correlation) was computed (see Table 8).

TABLE 8

VALUES OF SPEARMAN'S RHO (RANK-ORDER CORRELATION) BETWEEN MAX(μ_c^2), μ_c^2, AND BASIC TEST STATISTICS

	$\bar{X}$	%V	S	KR-21
max(μ_c^2)	.48	.86	.79	.88
$\overline{\mu_c^2}$	.29	.98	.80	.93

The results show that test mean had little relation to μ_c^2 (except as is discussed later), whereas the mean μ_c^2 was very highly correlated with both percent variance and KR-21. There was also a strong positive correlation between $\max(\mu_c^2)$ and both KR-21 and percent variance. That is, the greater the variance (or KR-21, or index of separation), the greater the maximum and average values of μ_c^2. These relationships are similar to those between basic test statistics and $\bar{\beta}$ or $\min(\beta)$, as reported earlier.

Because of the smoothness of the curves of Figures 38 through 45, attempts were made to find an algebraic function to describe the relationship between μ_c^2 and the test statistics. Several regression equations involving quadratic terms were tried, with the independent variables of test mean, percent mastery at the test mean, index of separation, percent mastery which produces the maximum value of μ_c^2, and both linear and binomial combinations of these. For more than two-thirds of these models, coefficients of determination were high, ranging from .84 to .93, but there was not enough consistency among regression coefficients to warrant any strong generalization. In summary, visual inspection of the family of curves provided just about as much information as these non-linear analyses of regression: there is a non-linear relationship between percent mastery and μ_c^2 (and other variables), but an algebraic expression of this relationship remains undiscovered.

In the earlier research cited above (Marshall, 1973), it was stated that for bimodal distributions, μ_c^2 seemed to be very highly correlated with percent mastery, and was related to test mean and percent mastery via a bivariate linear regression equation. Figures 43 and 45 help explain the inconsistency between that conclusion and the conclusion presented here. The earlier research used criterion levels of .6 and higher only, corresponding roughly to the left halves of these graphs. It is now evident that the erroneous conclusion of linearity was reached using such incomplete and unrepresentative data. The earlier report also asserted that the linear relationship was less strong for unimodal distributions, such as that represented by Figure 44. The relationship is clearly non-linear in the left half of that graph.

μ_c^2 and the number of examinees

For a given set of test parameters and a given criterion level, variation in the number of examinees did not seem to affect the value of μ_c^2. This was expected since μ_c^2 is the ratio of sums of squares, and hence increasing the number of examinees should affect both terms of the ratio equally.

Figure 46 shows a scatterplot of values of μ_c^2 calculated on 2N (or, as before, 4N) examinees against μ_c^2 calculated on N examinees.

Regression analysis showed the linear correlation of the pairs of values to be very high, .981. The obtained regression equation was $\hat{\mu}_c^2(2N) = -.004781 + .9931\ \mu_c^2(N)$, close enough to the model $\mu_c^2(2N) = \mu_c^2(N)$ to warrant its acceptance as the model that obtains in the population.

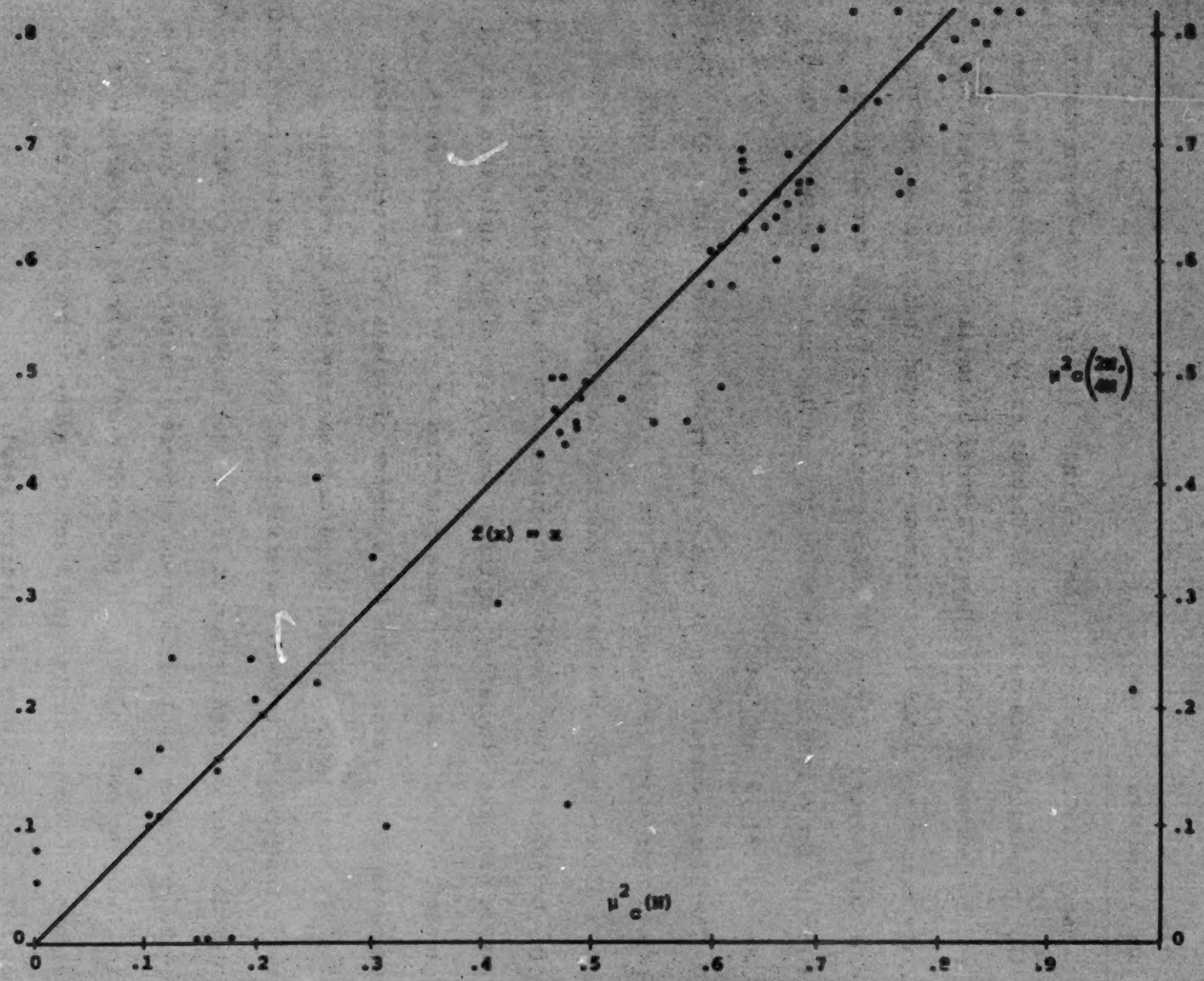


Figure 46. Scatterplot of μ_c^2 for 2N (or 4N) examinees against μ_c^2 for N examinees.

μ_c^2 and the number of items

Harris (1972a) indicates that his index is for "fixed-length mastery tests," presumably because there is theoretically no interaction between μ_c^2 and the number of items. Harris's index is unlike the classical reliability measures and the two criterion-referenced indices discussed thus far in this regard. Figure 47 shows a scatterplot of μ_c^2 for 2n items against μ_c^2 for n items, with n = 10, 20, and 40; and for criterion levels of .6, .7, .8, .9, and 1.0.

The linear correlation of this scatterplot was very high, .979. The obtained regression equation was $\hat{\mu}_c^2(2n) = -.0721 + 1.075\ \mu_c^2(n)$. This appears different enough from the expected no-effect model of $\mu_c^2(2n) = \mu_c^2(n)$ to suggest that another model might be more appropriate, but experience with the (simulated) empirical properties of μ_c^2 indicates another explanation. The data points were generated at five criterion levels, enumerated above, rather than for a number of values of percent mastery; yet μ_c^2 is more closely related to percent mastery than to criterion level. Depending on the score distribution, the percent mastery can fluctuate greatly for a given criterion level. For example, in the data discussed here, a criterion level of .8 produced percent mastery values ranging from 0 to .81. The model $\mu_c^2(2n) = \mu_c^2(n)$ would more likely be appropriate if the data had been generated for a set of values of percent mastery rather than for a set of values of criterion level.

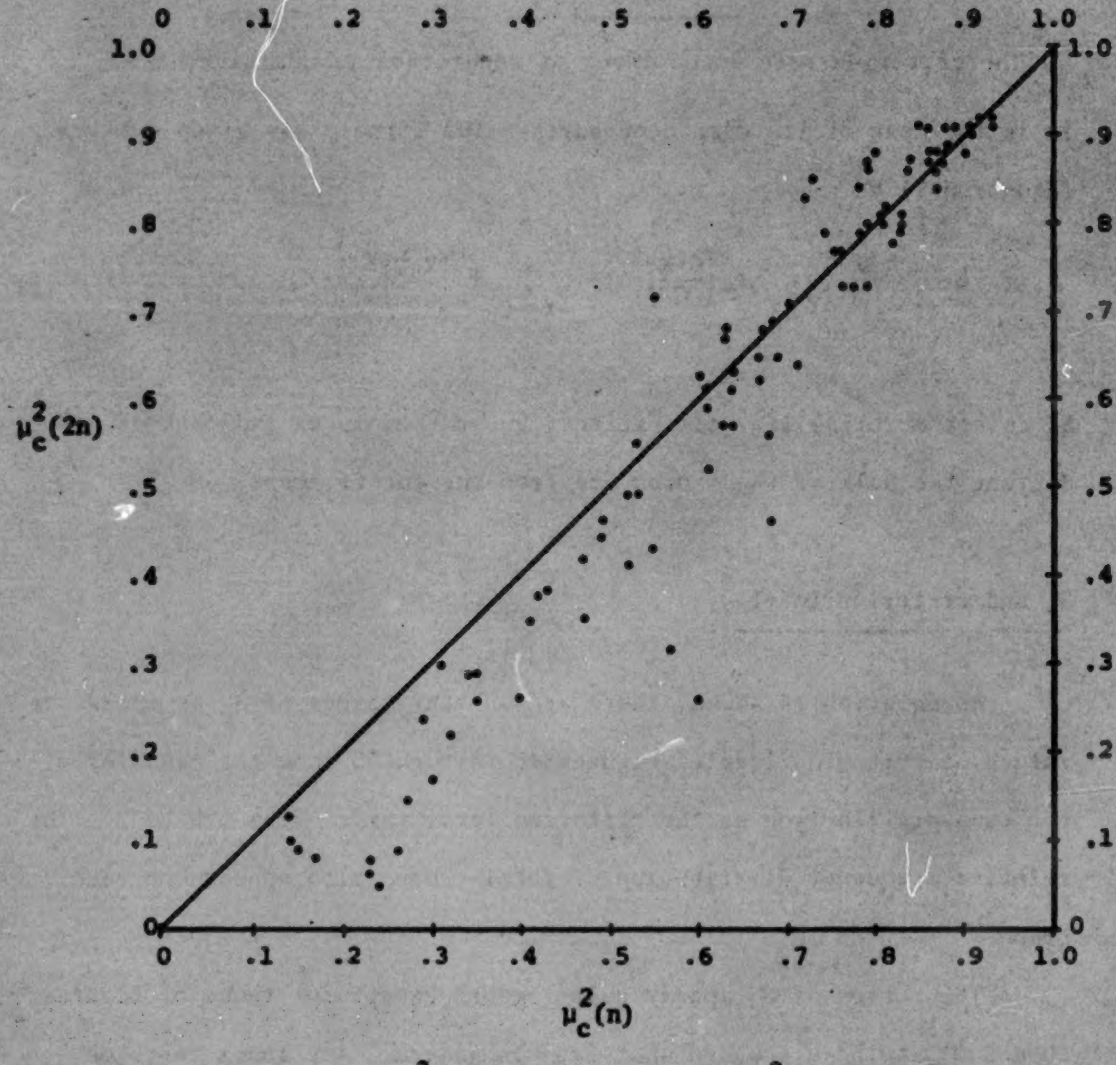


Figure 47. Scatterplot of μ_c^2 for 2n items against μ_c^2 for n items.

Characteristics of S_c

The criterion-referenced index of separation is additive, i.e., it is the mean of its component parts. The formula was given earlier (Equation 7) as

$$S_c = \frac{1}{N}\left(\sum_{X \leq C} f_x \left(\frac{C-X}{C}\right)^2 + \sum_{X > C} f_x \left(\frac{X-C}{n-C}\right)^2 \right) \qquad [14]$$

S_c is not a reliability coefficient, but rather is an indicant of how distant the bulk of the scores are from the cutoff score.

S_c and criterion level

As Equation 14 shows, there are as many values of S_c as there are values of criterion level. Figures 48 through 55 show the behavior of S_c for each distribution as the criterion level varies from .05 to 1. The relative frequency distribution of total scores also appears on each graph.

The curves of S_c appear quite smooth except for those of Figures 50 and 51, to be discussed shortly. In general, the index takes on lower values than do the other indices reported herein. There appears to be no tendency for S_c to approach either 0 or 1 as criterion level approaches 0 or 1.

Fig. 48

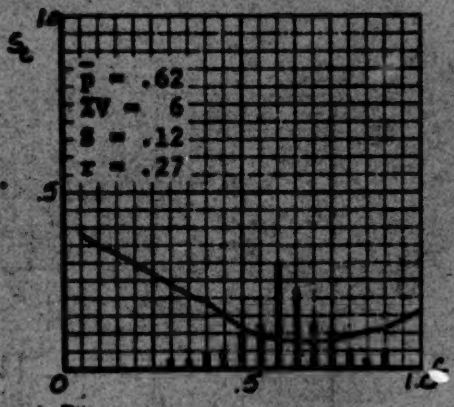


Fig. 49

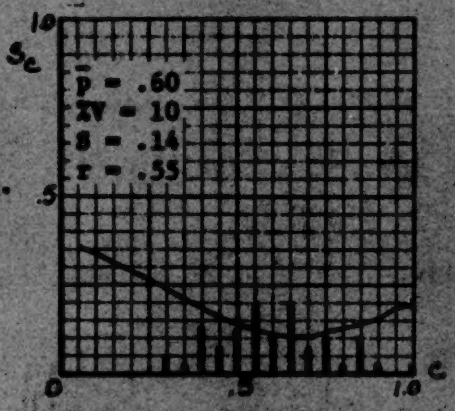


Fig. 50

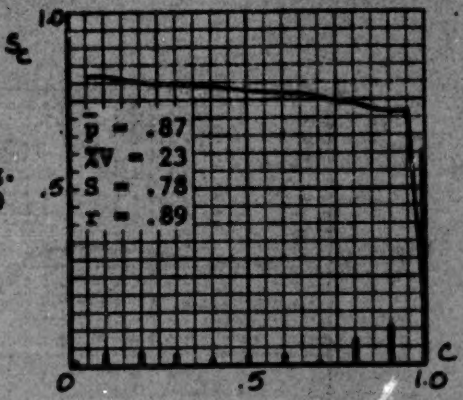
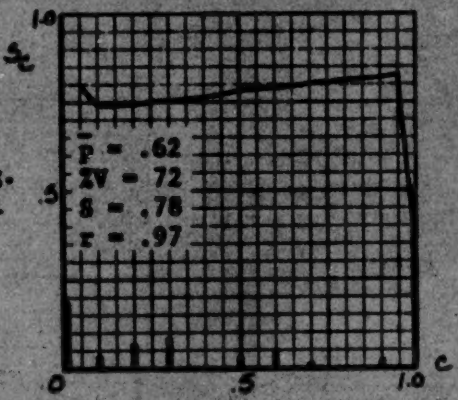


Fig. 51



Figures 48-51: Graphs of S_c against criterion level, with score distribution relative frequencies, for parameter sets 1-4.

Fig. 52

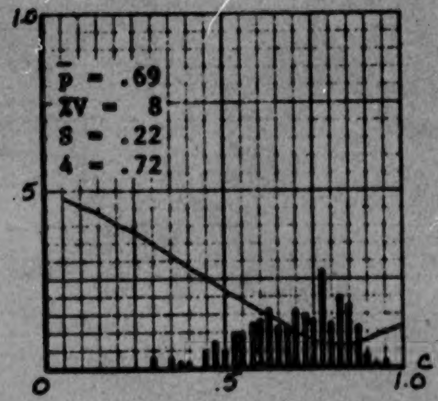


Fig. 53

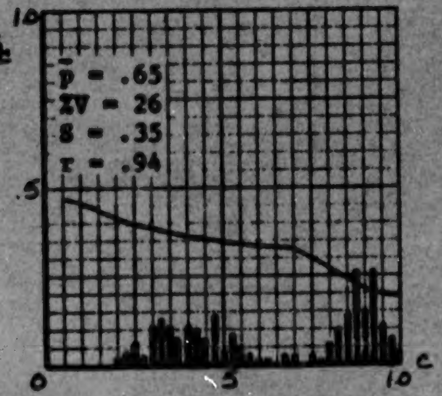


Fig. 54

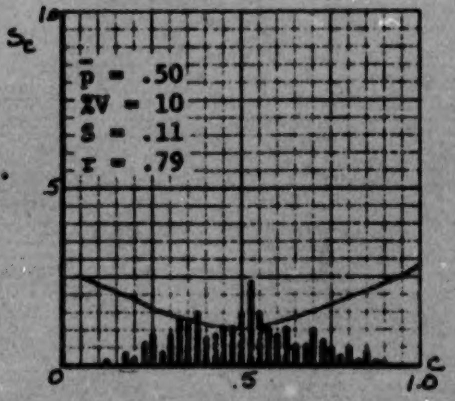
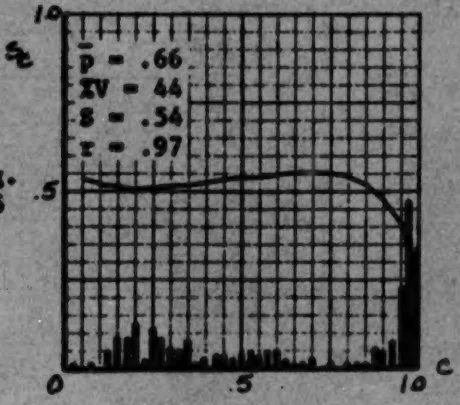


Fig. 55



Figures 52-55: Graphs of S_c against criterion level, with score distribution relative frequencies for parameter sets 5-8.

S_c and the score distribution

S_c seems to reflect the mode(s) of the score distribution, as does coefficient beta, but not always in the same way. This is particularly evident for extremely skewed or J-shaped distributions, such as are represented by Figures 50 and 51. On those graphs, the value of S_c drops sharply to correspond with the equally sharp mode at $X = n$.

S_c and basic test statistics

The size of (but not the variance in) the index appears to depend on the location of the test mean: the farther away the test mean (expressed as a percent) is from .5, the higher the overall value of the index until (as in Figures 50 and 51) the criterion corresponds to the mode. This appears to be the only consistent relationship between S_c and basic test statistics.

S_c and the number of examinees

For a given set of test parameters and a given criterion level, variation in the number of examinees did not seem to affect the value of S_c. This is reasonable in light of Equation 14, in which the effects of increasing the number of examinees should cancel out algebraically. Figure 56 shows a scatterplot of values of S_c calculated for 2N or 4N examinees against S_c calculated for N examinees, as was done for the other indices.

Regression analysis showed the linear correlation of this scatterplot to be unusually high, .997. The obtained regression equation was

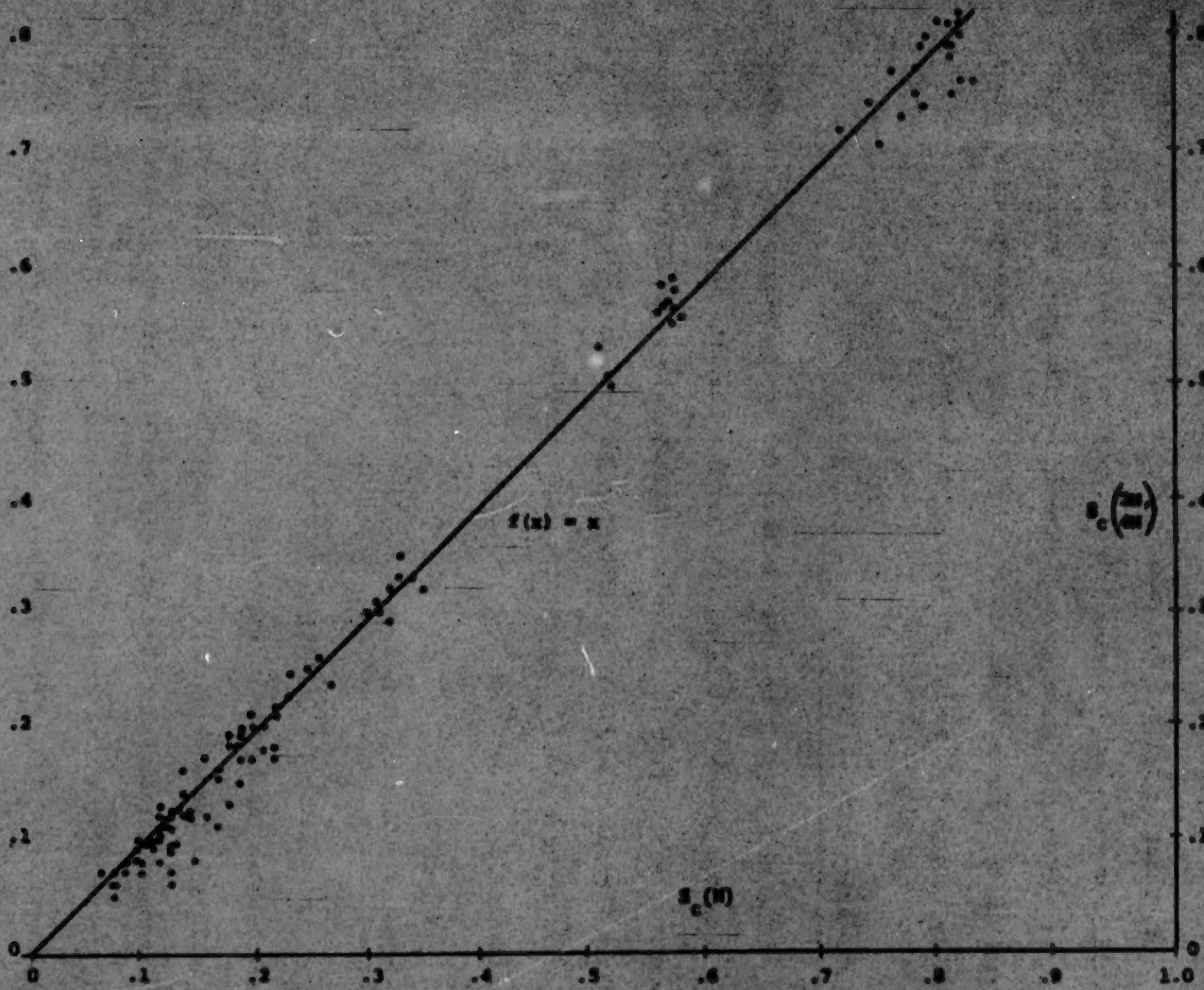


Figure 56: Scatterplot of S_c for 2N (or 4N) examinees against S_c for N examinees.

132

133

$\hat{S}_c(2N) = -.006453 + 1.009\ S_c(N)$, quite close to the model $S_c(2N) = S_c(N)$. Thus S_c is not affected by variation in the number of examinees.

S_c and the number of items

Figure 57 is a scatterplot of S_c for 2n items plotted again S_c for n items, with n and criterion levels as before.

Figure 57 shows that the points hew to the linear model. Regression analysis yielded a very high correlation of .997, and a regression equation of $\hat{S}_c(2n) = -.01669 + 1.003\ S_c(n)$, very close to the model $S_c(2n) = S_c(n)$. Hence S_c, unlike certain other indices, is apparently not affected by variation in the number of items.

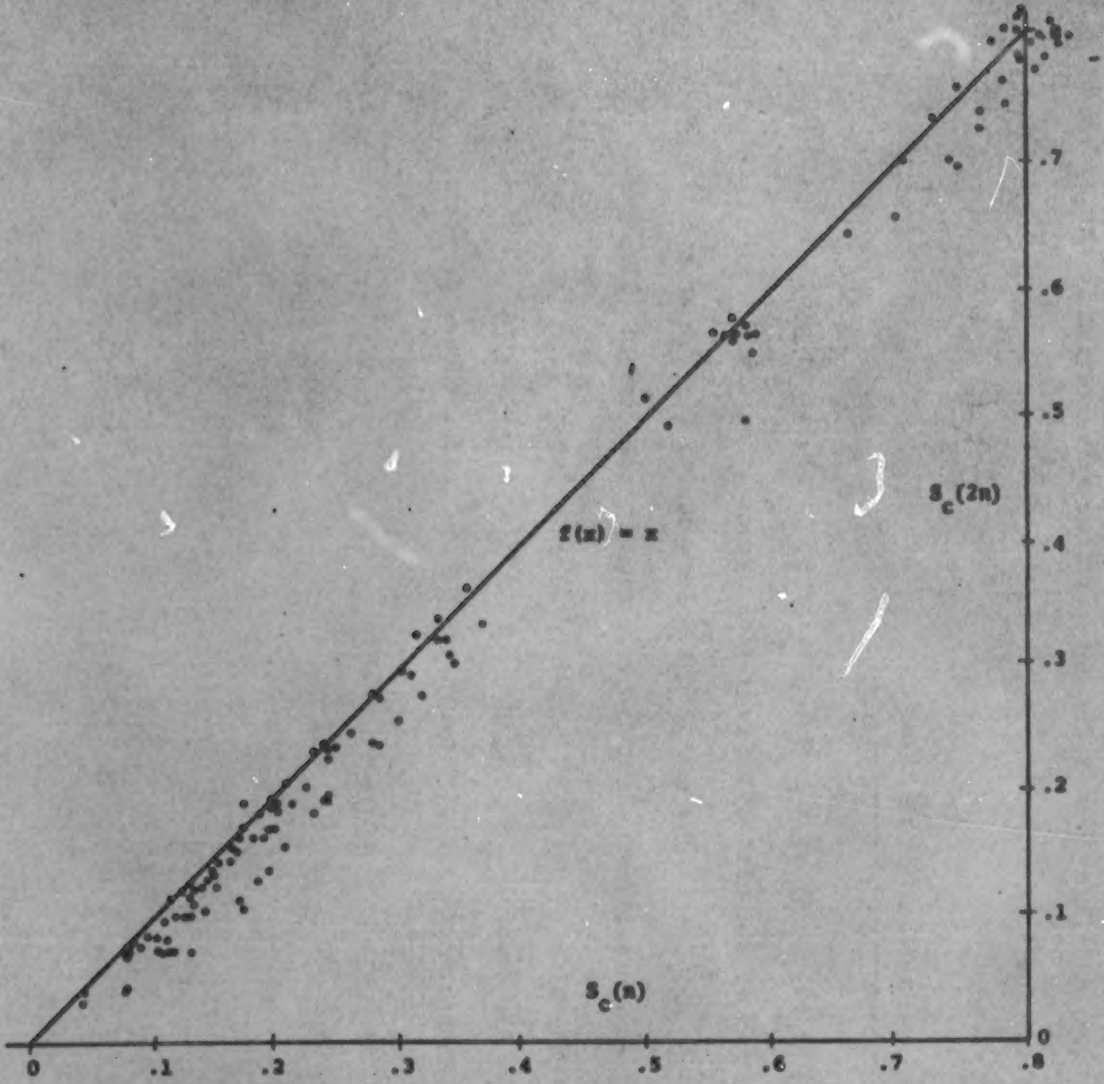


Figure 57: Scatterplot of S_c for 2n items against S_c for n items.

Relations Among Criterion-Dependent Indices

Two other indices enter into the analysis at this point: The cosine-pi estimate (r_{cospi}) of the tetrachoric correlation coefficient and the phi coefficient (r_ϕ). These indices are calculated from the "grand" fourfold table resulting from all possible split-half categorizations described near the end of Chapter IV, under which conditions r_ϕ is identical to coefficient kappa. All three indices were defined and briefly discussed in Chapter IV.

One way to summarize much of the data is to superimpose, for each parameter set, the individual graphs of the four indices presented earlier plus two more (but note that μ_c^2 is now plotted against criterion level rather than percent mastery). Figures 58 through 65 show values of β, k^2_{TX}, μ_c^2, S_c, r_{cospi} and r_ϕ, as well as the relative frequency distributions of total scores, for each of the eight parameter sets, using criterion level as the independent variable. In many of the graphs, it appears that these six indices are roughly grouped into three families: β, k^2_{TX} and S_c in one, r_{cospi} and r_ϕ in another, and (with some exceptions) μ_c^2 by itself. More will be said about these apparent interrelationships later.

Notice that r_{cospi} runs off the lower edge of most graphs at the extreme criterion levels. This is due to the occurrence of an empty cell in one of the diagonals of the fourfold table used in computing r_{cospi} by the formula given earlier as Equation 8. When one of these diagonal cells is empty, as is often the case at extremely low or high criterion levels, r_{cospi} is -1, even though the coefficient may have

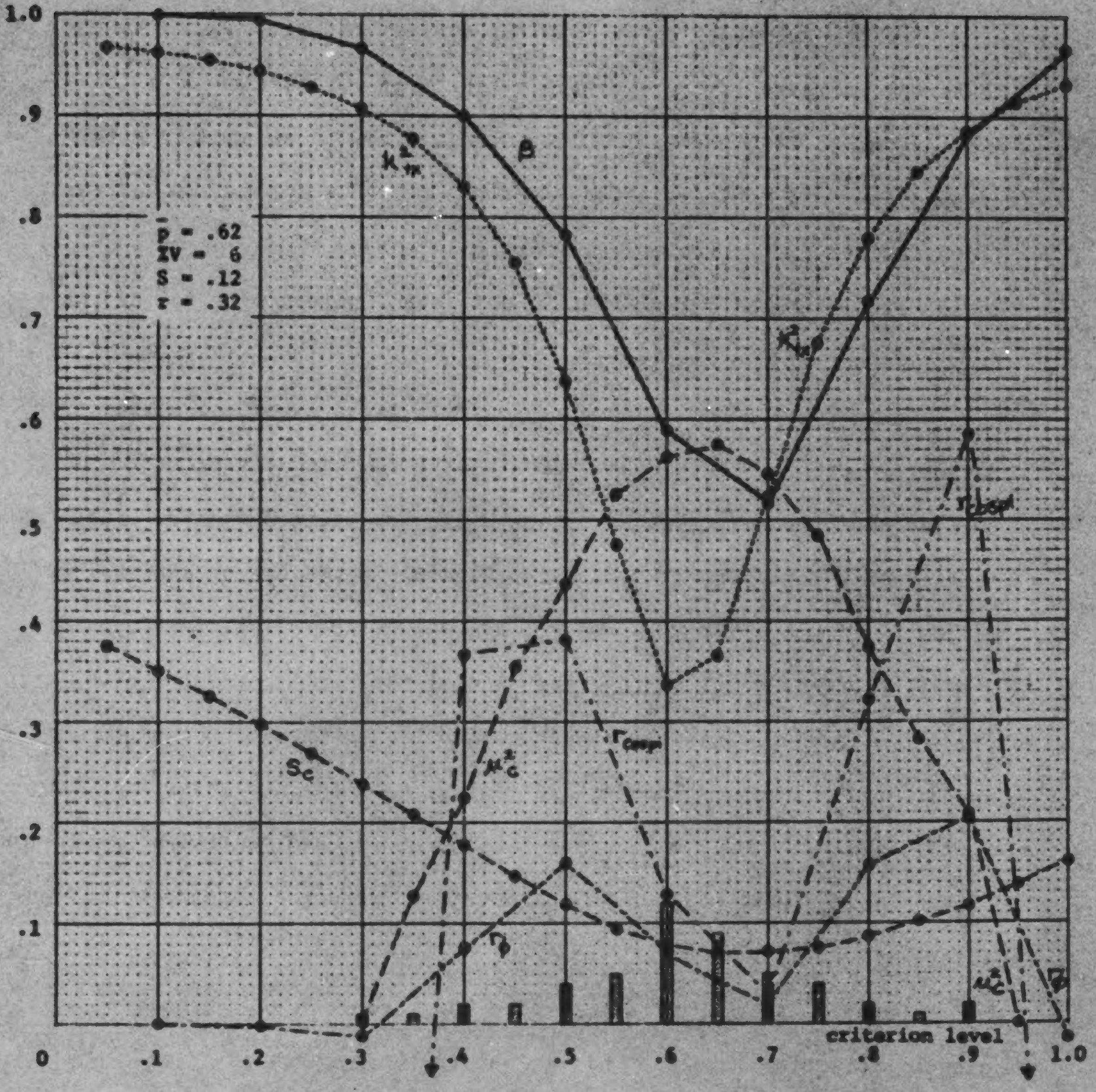


Figure 58. Indices vs. criterion level; parameter set 1.

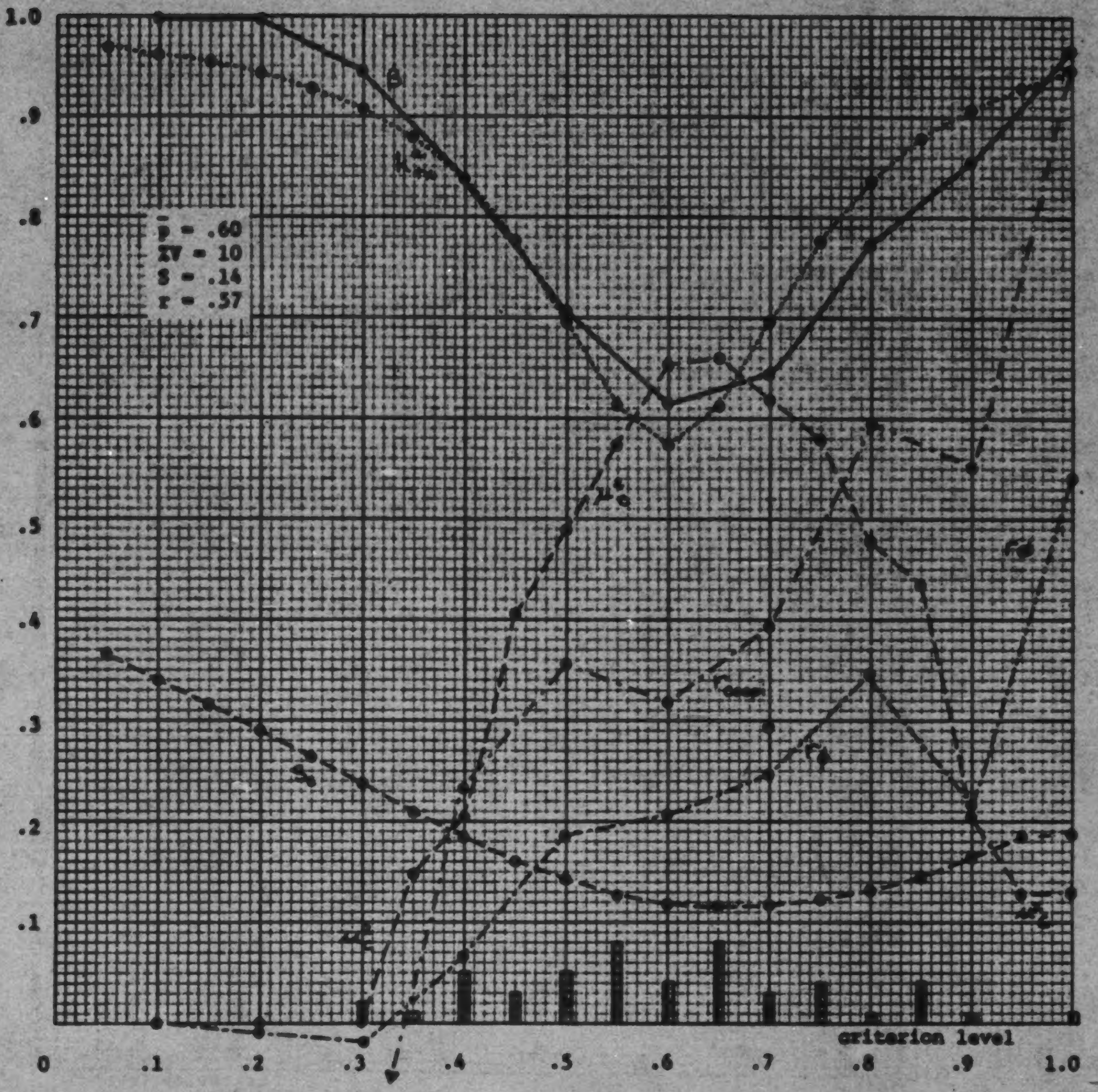


Figure 59. Indices vs. criterion level; parameter set 2.

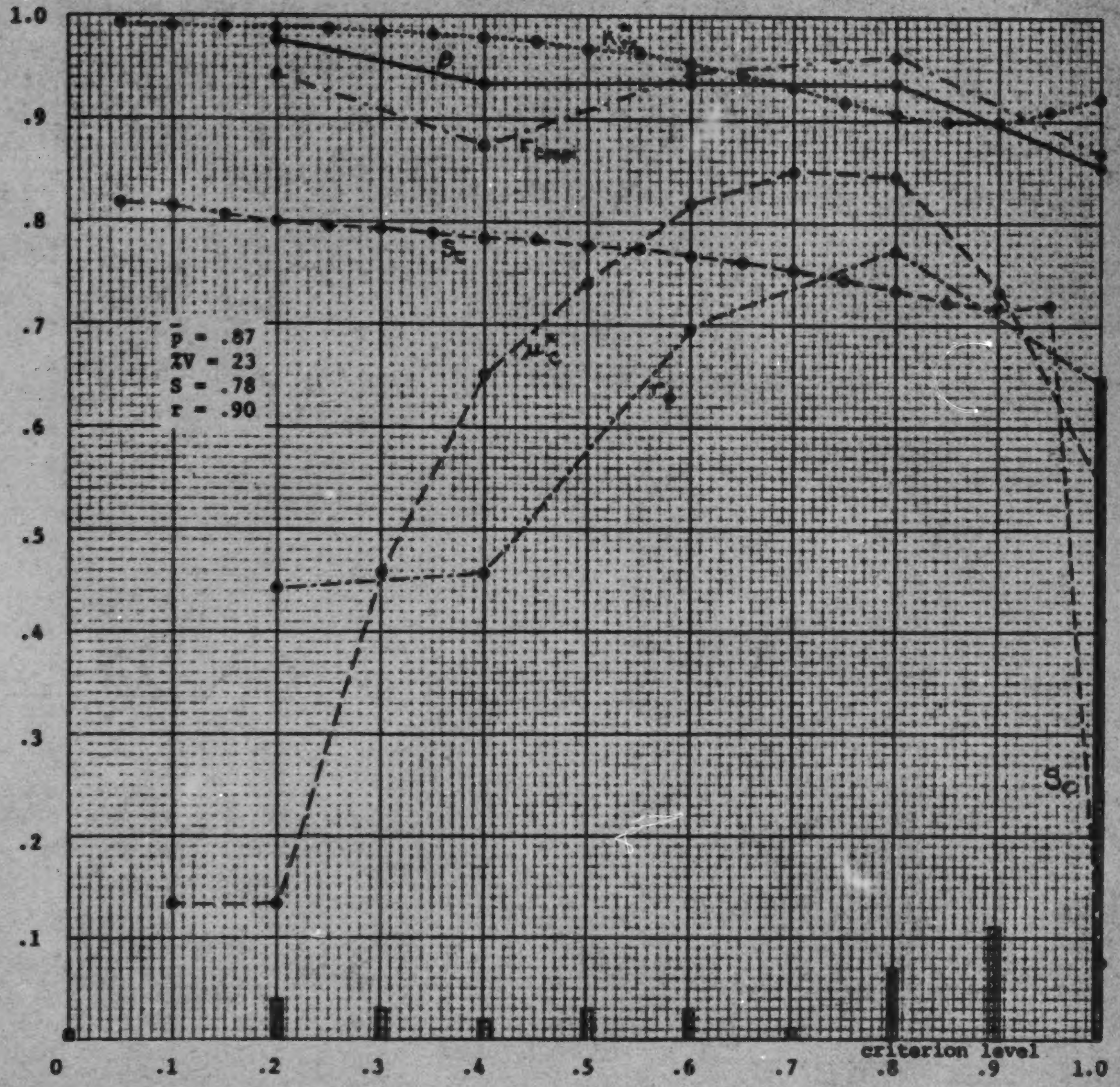


Figure 60. Indices vs. criterion level; parameter set 3.

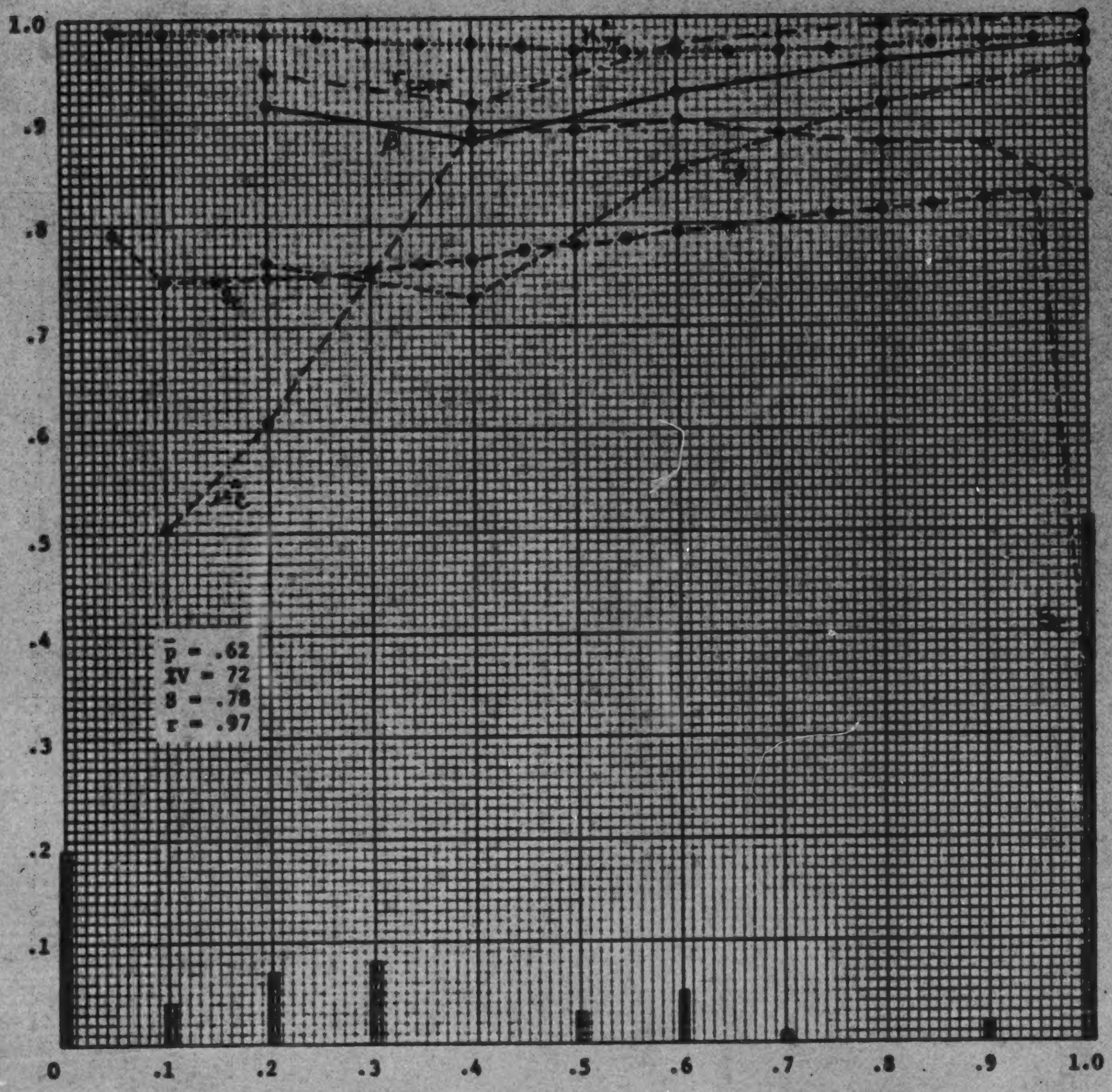


Figure 61. Indices vs. criterion level; parameter set 4.

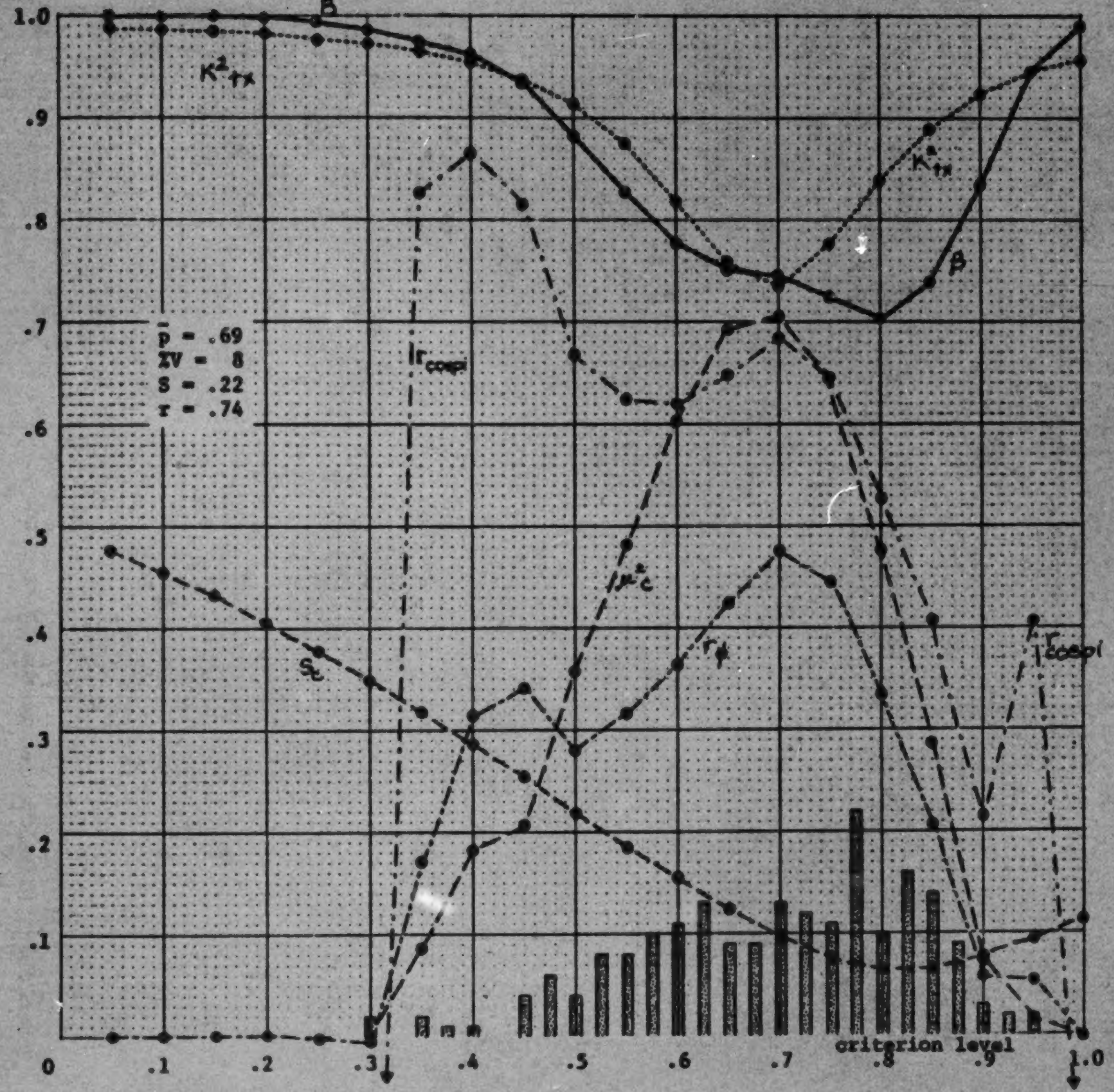


Figure 62. Indices vs. criterion level; parameter set 5.

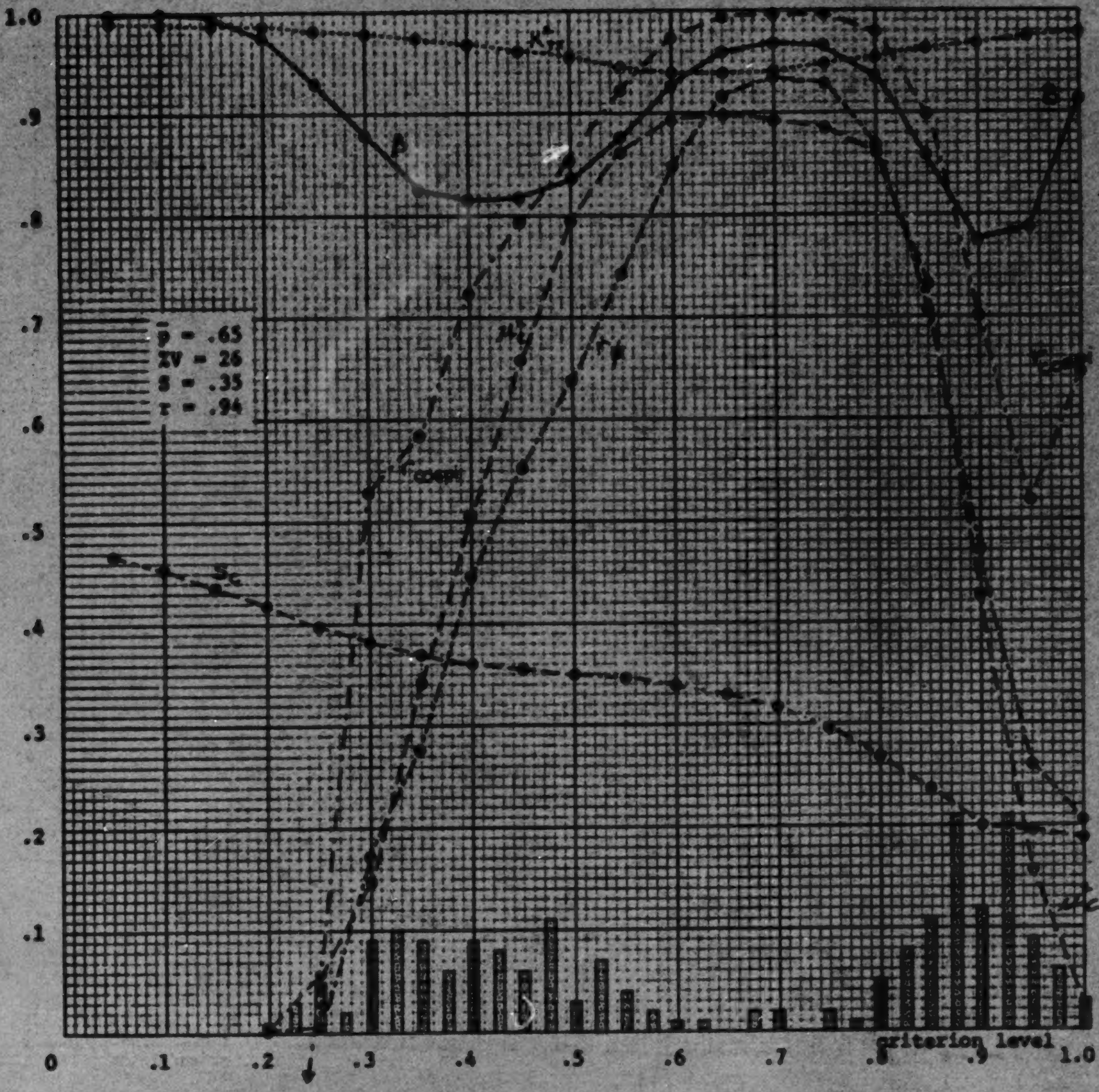


Figure 63. Indices vs. criterion level; parameter set 6.

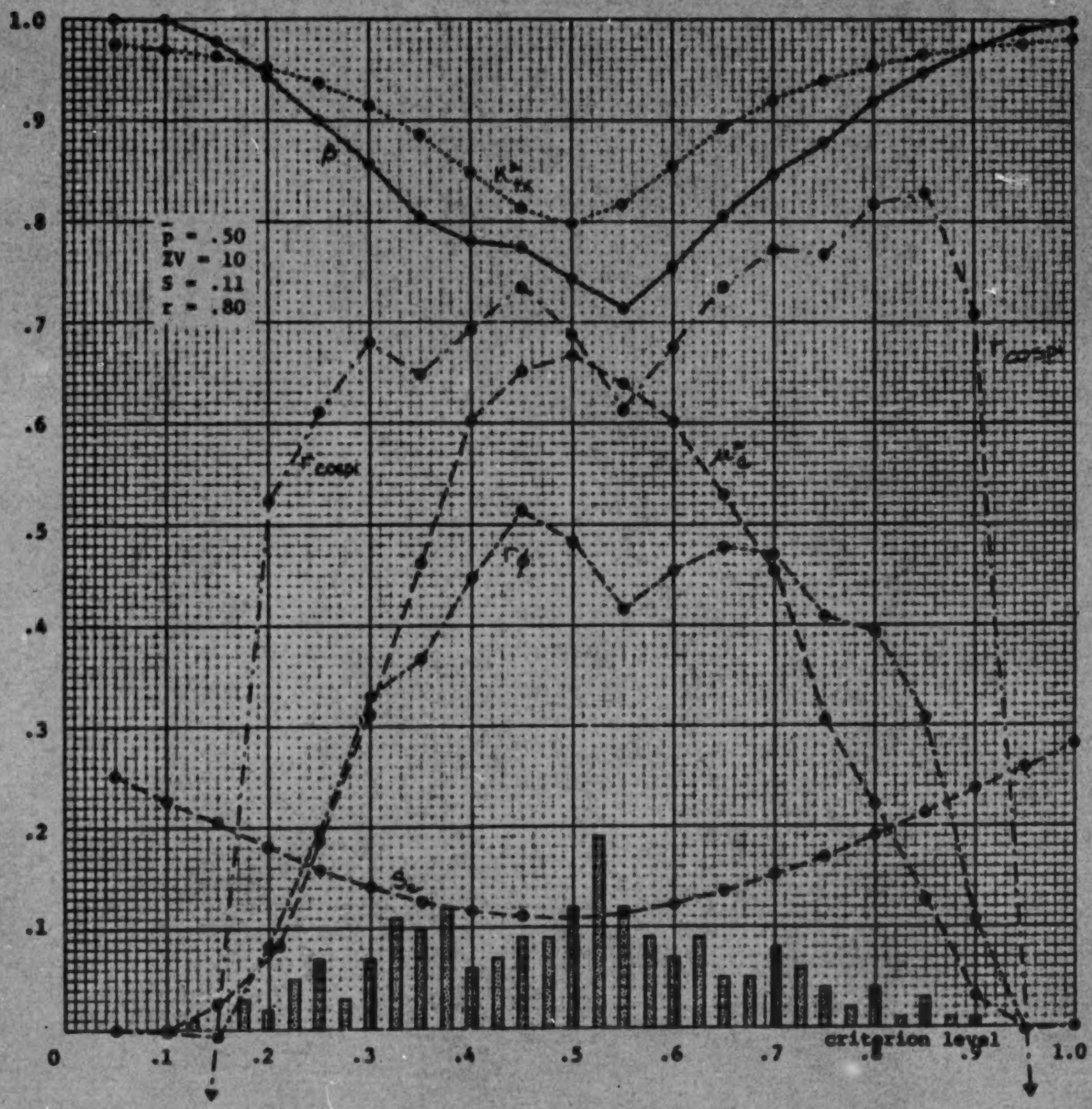


Figure 64. Indices vs. criterion level, parameter set 7.

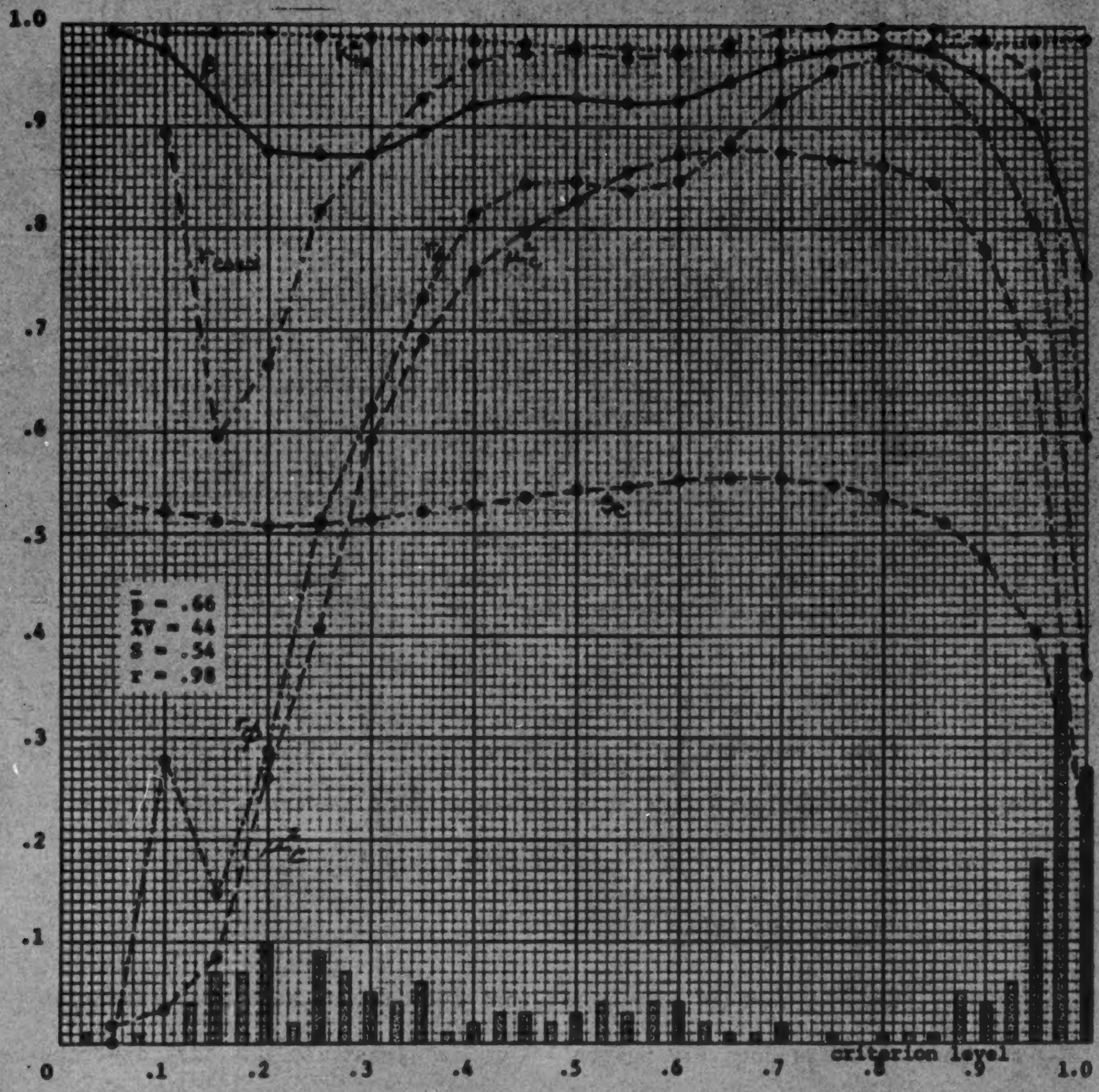


Figure 65. Indices vs. criterion level; parameter set 8.

quite a different value when the cell is nearly empty. For example, Table 9 shows, for the score distribution corresponding to Figure 64, the proportions within the four cells and the value of r_{cospi} for criterion levels of .90 and .95.

TABLE 9

EXTREME FLUCTUATIONS IN r_{cospi}

Criterion Level	Proportion in Cell				r_{cospi}
	A	B	C	D	
.90	.0020	.0145	.0145	.9690	.7097
.95	.0000	.0047	.0047	.9907	-1.0000

Because of this property of r_{cospi}, some of the analyses that follow might have been substantially altered if these extreme and unrepresentative values had been rescored or excluded from the data.

Coefficient beta and other indices

Figures 58 through 65 suggest that coefficient beta measures much the same thing as does Livingston's k^2_{TX}, at least for unimodal distributions. The two indices appear to have similar fluctuations as the criterion level varies, and they are generally close in value at each criterion level. The major difference is that β is sensitive to (has minima near) the mode(s) of the distribution, whereas k^2_{TX} is sensitive to (has minimum at) the mean of scores. Where the mean and mode more or less coincide, as in Figures 59 and 64, the coefficients are almost equal in value. For a bimodal distribution such as in Figures 63 or 65, however, the difference between them is clear. Since a true CRT could well be expected to have a bimodal distribution, this difference between the two coefficients is important.

Stepwise regression analyses bear out these intuitive arguments (see Appendix D for tables of data). In the regression model with β as the dependent variable (Table D-1) and with test mean, percent variance, KR-21, criterion level, k^2_{TX} and S_c as the independent variables, k^2_{TX} was always first to enter the regression equation (and hence would be closest and most influential in a "statistical sociogram") for each of the five unimodal distributions, and accounted for between 71% and 92% of the variance in β. Also consistent with the intuitive arguments, the amount of variance accounted for was 71% and 83% for the two distributions in which the mean and mode were some distance apart, and was higher for the distributions in which they more nearly coincided. The regression coefficient was always positive and with but one exception lay between .64 and .90. For each bimodal distribution, k^2_{TX} always entered the regression equation also but was never the first variable to do so, and it accounted for very little variance in β.

When all unimodal distributions were taken as a group, k^2_{TX} was again the first variable in the equation and accounted for all but 6% of the variance explained by that model; it did not even enter the equation when all bimodal distributions were taken as a group.

From the above data, which are rather consistent for stepwise regression analyses, it seems reasonable to conclude the following: for a unimodal test, β and k^2_{TX} measure much the same thing and result in similar values, but this relationship is weaker when the mean and the mode are not proximate; for bimodal tests, the two indices are sensitive to different properties of the score distribution.

Coefficient beta also has a moderately strong relationship with S_c. In the regression analysis discussed above, S_c also always entered the regression equation. For each unimodal distribution it was always the second variable to enter; for two of the three bimodal distributions (Figures 63 and 65), it was the first variable to enter, but accounted for only 29% and 52% of the variance of β. Figures 58 through 65 show that the curve of S_c over criterion levels did not generally fluctuate as much as did the curve of β, and S_c generally has a much lower value than does β. Nonetheless, they seem to measure somewhat similar things.

When r_{cospi} and r_{ϕ} were allowed to enter the regression equation, the results were not consistent. In one instance (Figure 61), r_{ϕ} was the first variable to enter and accounted for 95% of the variance, but this was a unique situation. Likewise, when all bimodal distributions were taken as a group, both r_{ϕ} and r_{cospi} entered the equation and together accounted for about half of the explained variance. However, this same pattern did not hold for individual bimodal distributions.

k^2_{TX} and other indices

When Livingston's k^2_{TX} was the dependent variable in the stepwise regression analysis (see Table D-2 in Appendix D), the results were less consistent than when β was the dependent variable. For instance, β did not always enter the equation when all variables were allowed to do so, even for unimodal distributions. However, it was the first variable to enter for four of the five unimodal distributions when the independent variables were restricted to the criterion-dependent test indices. Also, as in the analysis of β, when all unimodal distributions were taken as a group, β was the first to enter and accounted for 83% of the variance of k^2_{TX}, no matter which variables were allowed to enter the equation. A similar result occurred when all distributions were taken as a group. When all bimodal distributions were taken as a group, β did not enter the regression equation. Thus it is clear that k^2_{TX} measures much the same thing as does β, particularly for unimodal distributions.

For most of the distributions, μ^2_c also entered the regression equation, but the regression coefficients and the amount of variance accounted for were inconsistent. For the three distributions for which μ^2_c was the first to enter the equation, (Figures 61, 62 and 65), between 69% and 92% of the variance in k^2_{TX} was accounted for by μ^2_c, and the regression coefficients were all negative. Also, when all bimodal distributions were taken as a group, and all criterion-dependent indices were allowed to enter the equation, μ^2_c (with a negative regression coefficient) accounted for 25% of the variance in k^2_{TX}. Nonetheless, there does not seem to be sufficient evidence to generalize.

μ_c^2 and other indices

In the stepwise analysis of regression with μ_c^2 as the dependent variable and the other criterion-dependent test indices as the independent variables (Table D-3), r_ϕ was the first to enter the equation for three of the distributions (Figures 60, 63 and 65.). For the other five distributions, either β or k^2_{TX} was the first variable to enter, and the regression coefficients were always negative. This is an indication that μ_c^2 measures something opposite to what β (or k^2_{TX}) measures. For each distribution, r_ϕ was always either the first or second variable to enter the equation, and the regression coefficient was always positive.

When unimodal, bimodal, and all distributions were taken as groups, r_ϕ was also the first entering variable, accounting for 61%, 94%, and 79% of the variance, respectively. Hence it seems clear that, particularly for bimodal distributions, μ_c^2 and r_ϕ measure similar things.

S_c and other indices

When all variables (basic test statistics and parameters, criterion level, percent mastery, and the criterion-dependent test indices) were the free variables in the analysis, the results for S_c were not consistent. However, when this set was restricted to the criterion-dependent test indices (see Table D-4), coefficient beta was the predominant variable for all but two distributions (Figures 59 and 61), suggesting that S_c is in some way associated with β (and therefore with k^2_{TX}). However, the percent of variance in S_c accounted for by β was not always high. Moreover, when unimodal, bimodal, and all distributions were taken as groups, the results were inconclusive.

CHAPTER VII

SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH

Summary

In Chapter I it was stated that an increased acceptance of the interrelated notions of behavioral objectives, individualized instruction, and mastery learning has given rise to new kinds of educational tests. One of these new kinds of tests has as its purpose the efficient separation of the sample of examinees into two groups, often labeled "nonmastery" and "mastery." When an examinee has only two courses of action available after taking this kind of test--stay in the instructional module covered by the test or go on to studying the next module--his "score" need only be reported in terms of this dichotomy. Further subdivision of the test score scale serves no purpose; the dichotomy is sufficient to allow a decision leading to action to be made. A test of this type, which uses several items drawn from a well-defined universe to measure a single, narrow behavioral objective, and whose results yield a dichotomous categorization with reference to a predetermined criterion level, has herein been called a criterion-referenced test (CRT).

In Chapter II, some of the psychometric implications of the differences between a CRT and the more familiar norm-referenced test (NRT) were given. It was shown that the purpose, desired score distributions, test specifications, construction, and use in decision-making of CRTs are not generally the same as for NRTs. It was also shown that the classical and generally accepted mathematical model and assumptions that underlie the definitions of traditional measurement error and NRT test reliability do not apply to the dichotomous decision-making facet of a CRT. Thus a new, dual mathematical true-score model for CRTs was proposed: a CRT has both a positional facet, concerned with the primal measuring process and consistent with the classical assumptions and the continuous true-score model of an NRT, and an operational facet, concerned with the dichotomous decision-making process and consistent with a Platonic (dichotomous) true-score model but not with the classical model. It was further argued that the meanings of reliability should be different for the two facets of a CRT. Whereas an NRT (or the positional facet of a CRT) is reliable insofar as an examinee receives the same score on two parallel sets of data, the operational facet of a CRT demands that the test must also be reliable insofar as the examinee receives the same dichotomous categorization from the two sets of data. But since a classical reliability estimate is inappropriate for this second facet of a CRT, what should take its place?

In Chapter III, an answer to this question is offered. An appropriate CRT reliability index ought to be founded on the notion of consistent categorizations. A single-administration coefficient that reflects this notion is the mean of all possible split-half coefficients of agreement, where the coefficient of agreement is the proportion of consistent categorizations, i.e., the proportion of entries in the main diagonal of a fourfold mastery/nonmastery contingency table. Such an index, labeled coefficient beta (β) because of the mean split-half analogy with

Cronbach's alpha, was derived, and theoretical and computational formulas were given. The computational adjustments required when the test has an odd number of items were noted. Certain technical characteristics of coefficient beta were mentioned, and β was shown to satisfy a list of CRT index criteria that were proposed in Chapter II. Finally, coefficient beta was extended to trichotomous data, and a formula for the modified coefficient was given.

In Chapter IV, three other recent criterion-dependent test indices were defined--$k^2_{T\bar{X}}$ (Livingston, 1972a), μ^2_c (Harris, 1972a), and S_c (introduced in the chapter) -- and their rationales were briefly discussed. Each index was tested against the CRT reliability index criteria proposed earlier. In addition, the cosine-pi estimate of the tetrachoric correlation coefficient and the phi coefficient were defined, and it was shown that either coefficient can be construed as a single-administration index if it is calculated from a fourfold table whose cells contain numbers resulting from all possible split-half mastery categorizations. It was shown that, under these conditions, the phi coefficient and Cohen's kappa coefficient are identical.

In Chapter V the questions investigated in the study were posed and the analytical methodology used to seek answers to them was discussed. The questions dealt with certain aspects of coefficient beta and the three other criterion-dependent indices: their characteristics, their interrelationships, their relationships to basic test statistics, and their behavior as criterion level changes and as the number of examinees and the number of items increases (and in the latter case, the degree to which the Spearman-Brown prophecy formula applies). The only feasible way to carry out this kind of study is with simulated data, and hence the computer program that generated the data for this study was described in this chapter. Included in this discussion were the equation used by the program to generate item-by-pupil response matrices, the available input parameters and output options, and the eight input parameter sets (and hence kinds of score distributions) that were selected for this study. The parameter sets were chosen to simulate three types of tests, discussed in the chapter.

In Chapter VI the results of the data generation were given in graphs and the data were analyzed through stepwise analyses of regression, both linear and non-linear. Characteristics of each of the four criterion-dependent test indices were given. For example, for all the score distribution types studied, consistently moderate to high correlations existed between the mean (over criterion level) of each of three of these indices and classical reliability (and in the case of μ_c^2, percent of maximum variance). None of the four criterion-dependent indices was affected by the number of examinees, which is reassuring. However, the indices varied in the degree to which they were affected by changes in the number of items. The criterion-referenced index of separation, S_c, and Harris's index of efficiency, μ_c^2, were not affected by the number of items, but β and k^2_{TX} were. The Spearman-Brown prophecy formula explained the behavior of k^2_{TX}, but the behavior of β was explained equally well by the Spearman-Brown prophecy model and the (linear) no-effect model. The empirical evidence showed that the variation in β as the number of items increased was best explained by a model that is an algebraic compromise between the Spearman-Brown and the no-effect models.

Other relationships were revealed. Perhaps most important and clear-cut among them was that for unimodal score distributions, coefficient beta seems to measure much the same thing as Livingston's k^2_{TX} -- their fluctuations over criterion level and their ranges of values were generally quite similar--but for bimodal distributions this relationship does not hold. The reason is that β is sensitive to (has minima near) the mode(s) of the score distribution, consistent with the proposal that a CRT reliability index should have higher values as the bulk of scores depart from the cutoff score, whereas k^2_{TX} is sensitive to (has minimum at) the test mean.

There were moderately consistent correlations (over score distribution types) between β and S_c, between k^2_{TX} and β, and between μ^2_c and r_ϕ. Put differently, coefficients β, k^2_{TX}, and S_c seem to measure similar test result attributes, as do μ^2_c and r_ϕ (and therefore κ). However, there is a basic difference between the first group (β, k^2_{TX}, and S_c) and the second group (μ^2_c and r_ϕ): the indices in the former group tend to have higher values (1, in the case of β and k^2_{TX}) at the extremes of criterion level, whereas the latter group tend toward 0 at these same extremes.

To choose a "best" reliability coefficient for the operational facet of a CRT, one must take into account its premises, rationale, and characteristics. Of coefficients β, k^2_{TX}, and μ^2_c, only β is sensitive to the test mode(s) as distinct from the mean. Thus if it is desired that a

CRT operational reliability index have higher values as scores depart from the cutoff, coefficient beta is the reliability index that should be used.

SUGGESTIONS FOR FURTHER RESEARCH

The following research suggestions are based on the results of this study:

1. Coefficient beta increases as the number of items increases, and it is the mean coefficient of agreement calculated on all possible halves of a test. These two facts may suggest that β is really a half-test index, and that its value should somehow be stepped up if it is to be applied to a whole test.

At least three basic approaches could be made to the stepping-up procedure. One approach would be to provide a formula that produces a whole-test coefficient as a function of the half-test coefficient, similar to the Spearman-Brown prophecy formula or to Equation 12 in Chapter VI. Another approach would be to calculate coefficient beta on a test of twice as many items as are ultimately intended to be used and then drop, selectively or randomly, half the items. A third approach would be to estimate, based on the obtained score distribution, what the score distribution would be on a test twice as long, and then calculate β from the score distribution so estimated. This last approach seems to hold promise, and further research results using either a regression, a Bayesian, or a binomial model to estimate the double-length score distribution could prove fruitful. (See Appendix E for binomial model approach.)

2. In Chapter II it was argued that operational reliability of a CRT must be concerned with accuracy of placement to categories, and that one useful definition of such reliability would be the proportion of classifications which are correct classifications (see Table 3). It was further suggested that a meaningful CRT reliability coefficient would be a statistic which estimates or is a lower bound to this proportion.

Although it is intuitively reasonable to suppose that coefficient beta is related to this proportion of classifications that are correct classifications, such a conclusion has not yet been proved mathematically and affords a topic for future research.

3. Coefficient alpha is equal to the mean split-half classical reliability coefficient. Coefficient beta is equal to the mean split-half coefficient of agreement. For a given total score distribution, α takes on different values for different item-by-examinee response matrices, and β takes on different values for different criterion levels. Preliminary research indicates that, for a given response matrix, the mean value of β (over criterion level) is often close to the computed coefficient alpha. It may be that, for a given distribution of total scores, there is some relation (upper or lower bound? algebraic function? equality?) between the mean value of α (over response matrices) and the mean value of β (over criterion levels). This possibility would be interesting to investigate.

4. It was pointed out at the end of Chapter IV (see also Appendix A) that when the off-diagonal cells in the fourfold table are equal, the phi coefficient (r_ϕ^*) and coefficient kappa (κ^*) are identical. It was then

hypothesized, based on a small sample of score distributions, that κ^* (and thus r_ϕ^*) is a generally close lower bound to $\bar{\kappa}$, the mean split-half kappa coefficient. If this conjecture can be proved, one could use r_ϕ^* (Equation 10) to obtain a close lower bound to $\bar{\kappa}$.

5. At the end of Chapter III, coefficient beta was extended to incorporate trichotomous data. It may be that the coefficient can be further extended to incorporate data utilizing four classifications, or possibly generalized to any number of classifications. Extrapolation from an analysis of the formulas for β and β_3 suggests, however, that for an n-item test, the maximum number of classifications is $\frac{n}{2} + 1$.

REFERENCES

American Association for the Advancement of Science Commission on Science Education. The psychological bases of science -- a process approach. Washington, D. C.: American Association for the Advancement of Science, 1965.

Baker, F. B. Origins of the item parameters X_{50} and β as a modern analysis technique. Journal of Educational Measurement , 1965, 2, 167-180.

Berger, R. J. A measure of reliability for criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, March, 1970.

Blatchford, C. H. Experimental steps to ascertain reliability of diagnostic tests in English as a second language. Unpublished doctoral dissertation, Columbia University, 1970.

Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1 (No. 2).

Brennan, R. L. The evaluation of mastery test items. Final Report, Project no. 2B118, National Center for Educational Research and Development, U.S. Department of Health, Education and Welfare, Washington, D. C., 1974.

Brennan, R. L. & Stolurow, L. M. An elementary decision process for the formative evaluation of an instructional system. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Carroll, J. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Carver, R. P. Special problems in measuring change with psychometric devices. In B. Baxter (Ed.), Evaluative Research: Strategies and Methods. Pittsburgh: American Institutes for Research, 1970.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 213-220.

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 292-334.

Cronbach, L. J. & Gleser, G. C. Psychological tests and personnel decisions. (2nd ed.) Urbana: University of Illinois Press, 1965.

Darlington, R. B. & Bishop, C. H. Increasing test validity by considering interitem correlations. Journal of Applied Psychology, 1966, 50, 322-330.

Davis, F. B. Item analysis in relation to educational and psychological testing. Psychological Bulletin, 1952, 49, 97-121.

Developing Mathematical Processes Staff. Resource Manual, Topics 1-40, for Developing Mathematical Processes. Chicago: Rand-McNally, 1974.

Donlon, T. F. Some needs for clearer terminology in criterion referenced testing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1974.

Evans, J. Behavioral objectives are no damn good. In Technology and innovation in education (prepared by the Aerospace Education Foundation). New York: Praeger, 1968.

Flanagan, J. C. A proposed procedure for increasing the efficiency of objective tests. Journal of Educational Psychology, 1937, 28, 17-21.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1965.

Gessel, J. Prescriptive mathematics inventory. Monterey, Cal.: CTB/McGraw-Hill, 1972.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R. & Cox, R.C. Criterion-referenced testing for the measurement of educational outcomes. In Weisberger, R.A. (Ed.), Instructional process and media innovation. Chicago: Rand-McNally, 1968.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 733-764.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classifications: II. Further discussions and references. Journal of the American Statistical Association, 1959, 54, 123-163.

Guilford, J. P. Fundamental statistics in psychology and education. (4th ed.) New York: McGraw-Hill, 1965.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. An index of efficiency for fixed-length mastery tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972. (a)

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29. (b)

Harris, M. L. & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Hsu, T-C. Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Klausmeier, H. J., Quilling, M. R., Sorenson, J. S., Way, R. S. & Glasrud, G. R. Individually guided education and the multi-unit elementary school: Guidelines for implementation. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1971.

Klein, D. F. & Cleary, T. A. Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 1967, 68, 77-80.

Klein, S. P. & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 26. Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973.

Kosecoff, J. B. & Klein, S. P. Instructional sensitivity statistics appropriate for objective-based test items. CSE Report No. 91. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation, 1974.

Kuder, G. F. & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Livingston, S. A. A criterion-referenced application of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.(a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, *9*, 31. (b)

Livingston, S. A. Reply to Shavelson, Block and Ravitch's "Criterion-referenced testing: Comments on reliability." *Journal of Educational Measurement*, 1972, *9*, 139-140. (c)

Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Marshall, J. L. Reliability indices for criterion-referenced tests: A study based on simulated data. Paper presented at the annual meeting of the National Council on Measurement in Education, New Oreleans, February, 1973.

Marshall, J. L. & Haertel, E. H. A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., March-April, 1975.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley: McCutchan, 1974.

Nitko, A. J. A model for criterion-referenced tests based on use. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Novick, M. R. & Lewis, C. Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, *32*, 1-13.

Otto, W. & Askov, E. *Rationale and guidelines* for the Wisconsin Design for Reading Skill Development (3rd ed.) Minneapolis: National Computer Systems, 1974.

Ozenne, D. G. *Toward an evaluative methodology for criterion-referenced measures: Test sensitivity*. CSE Report No. 72. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation, 1971.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham, (Ed.), *Criterion-referenced measurement*. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rim, E-D. Livingston's reliability coefficient and Harris' index of efficiency: An empirical study of the two reliability coefficients for criterion-referenced tests. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April, 1974.

Raju, N. S. A note on Livingston's reliability for criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, February, 1973.

Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1974.

Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 1939, 9, 99-103.

Shavelson, R., Block, J., & Ravitch, M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Simon, G. B. Comments on "Implications of criterion-referenced measurement." Journal of Educational Measurement, 1969, 6, 259-260.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.

STEPREG1: Stepwise linear regression analysis. Madison: University of Wisconsin Academic Computing Center, 1973.

Sutcliffe, J. P. A probability model for errors of classification: I. General considerations. Psychometrika, 1965, 30, 73-96.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Wedman, I. Reliability, validity and discrimination measures for criterion-referenced tests. Educational Reports, Umeå (Umeå University, Sweden), 1973, Whole No. 4.

APPENDIX A

Supplementary Algebraic Derivations

A-1 Proof that $\sum_{X=0}^{n-1} g_x = \sum_{X=0}^{n} f_x$ if $g_x = \frac{n-x}{n} f_x + \frac{x+1}{n} f_{x+1}$

A-2 Relation between the two indices of separation

A-3 Equivalence of phi coefficient (r_ϕ^*) and coefficient kappa (κ^*) when off-diagonal cells are equal (B=C=E)

A-1 Proof that $\sum_{x=0}^{n-1} g_x = \sum_{x=0}^{n} f_x$, if $g_x = \frac{n-x}{n} f_x + \frac{x+1}{n} f_{x+1}$

$$\sum_{x=0}^{n-1} g_x = g_o + \sum_{x=1}^{n-1} g_x$$

$$= \left(\frac{n}{n} f_o + \frac{1}{n} f_1\right) + \left(\sum_{x=1}^{n-1} \frac{n-x}{n} f_x + \sum_{x=1}^{n-1} \frac{x+1}{n} f_{x+1}\right)$$

$$= f_o + \frac{1}{n} f_1 + \left(\sum_{x=1}^{n-1} \frac{n}{n} f_x - \sum_{x=1}^{n-1} \frac{x}{n} f_x\right) + \sum_{x=2}^{n} \frac{x}{n} f_x$$

where, in the last term, x was replaced by $x-1$

$$= f_o + \sum_{x=1}^{n-1} f_x + \frac{1}{n} f_1 - \left(\frac{1}{n} f_1 + \sum_{x=2}^{n-1} \frac{x}{n} f_x\right) + \left(\sum_{x=2}^{n-1} \frac{x}{n} f_x + \frac{n}{n} f_n\right)$$

$$= \sum_{x=0}^{n-1} f_x + f_n$$

$$= \sum_{x=0}^{n} f_x$$

A-2 Relation between the two indices of separation

In Chapter IV, the index of separation of total scores, S, is given (Equation 6) as:

$$S = 1 - \frac{4}{nN} \sum_{p} (X_p - \frac{1}{n} X^2_p) \text{ , where}$$

n = the number of items

N = number of persons, and

X_p = pth person's total score.

In addition, the criterion-referenced index of separation of total scores, S_c, is given (Equation 7) as:

$$S_c = \frac{1}{N} \left[\sum_{X \leq C} f_x \left(\frac{C-X}{C} \right)^2 + \sum_{X > C} f_x \left(\frac{X-C}{n-C} \right)^2 \right] \text{ , where}$$

n, N, and X_p are as above, f_x is the frequency of score X in the distribution of scores, and C is the criterion cut-off score. S can be shown to be a special case of S_c. If we start with the formulation of S_c and substitute $\frac{n}{2}$ for C, we obtain

$$S_c = \frac{1}{N} \left[\sum_{X \leq \frac{n}{2}} f_x \left[\frac{\frac{n}{2} - X}{\frac{n}{2}} \right]^2 + \sum_{X > \frac{n}{2}} f_x \left[\frac{X - \frac{n}{2}}{n - \frac{n}{2}} \right]^2 \right]$$

$$= \frac{1}{N} \sum_{X} f_x \left[\frac{\frac{n}{2} - X}{\frac{n}{2}} \right]^2$$

$$= \frac{1}{N} \sum_{p} \left[\frac{\frac{n}{2} - X_p}{\frac{n}{2}} \right]^2$$

$$= \frac{1}{N} \sum_{p} \left[1 - \frac{2 X_p}{n} \right]^2$$

$$= \frac{1}{N} \sum_{p} \left[1 - \frac{4}{n} X_p + \frac{4}{n^2} X_p^2 \right]$$

$$= 1 - \frac{4}{nN} \sum_{p} \left[X_p - \frac{1}{n} X_p^2 \right]$$

$$= S$$

A-3 Equivalence of the phi coefficient ($r_\phi{}^*$) and coefficient kappa (κ^*) when off-diagonal cells are equal (B=C=E)

$$\kappa^* = \frac{\frac{A}{N} + \frac{D}{N} - \left[\frac{(A+E)(A+E)}{N^2} + \frac{(D+E)(D+E)}{N^2}\right]}{1 - \left[\frac{(A+E)(A+E)}{N^2} + \frac{(D+E)(D+E)}{N^2}\right]}$$

$$= \frac{AN + DN - (A+E)^2 - (D+E)^2}{N^2 - (A+E)^2 - (D+E)^2}$$

$$= \frac{A(A+D+2E) + D(A+D+2E) - (A+E)^2 - (D+E)^2}{(A+D+2E)^2 - (A+E)^2 - (D+E)^2}$$

$$= \frac{A^2 + AD + 2AE + AD + D^2 + 2DE - A^2 - 2AE - E^2 - D^2 - 2DE - E^2}{A^2 + D^2 + 4E^2 + 2AD + 4AE + 4DE - A^2 - 2AE - E^2 - D^2 - 2DE - E^2}$$

$$= \frac{2AD - 2E^2}{2AD + 2AE + 2DE + 2E^2}$$

$$= \frac{2(AD-E^2)}{2(A+E)(D+E)}$$

$$= \frac{AD-E^2}{(A+E)(D+E)}$$

$$= r_\phi{}^*$$

APPENDIX B

Graphs of $\phi(X)$ for each Score X, for Selected Criterion Levels and Number of Items

B-1 $\phi(X)$ for each X on a 4-item test for all meaningful criterion levels.

B-2 $\phi(X)$ for each X on an 8-item test for three selected criterion levels.

B-3 $\phi(X)$ for each X on a 16-item test for four selected criterion levels.

B-4 $\phi(X)$ for each X on a 32-item test for seven selected criterion levels.

B-5 $\phi(X)$ for each X on a 10-item test for three selected criterion levels.

B-6 $\phi(X)$ for each X on a 20-item test for five selected criterion levels.

B-7 $\phi(X)$ for each X on a 40-item test for five selected criterion levels.

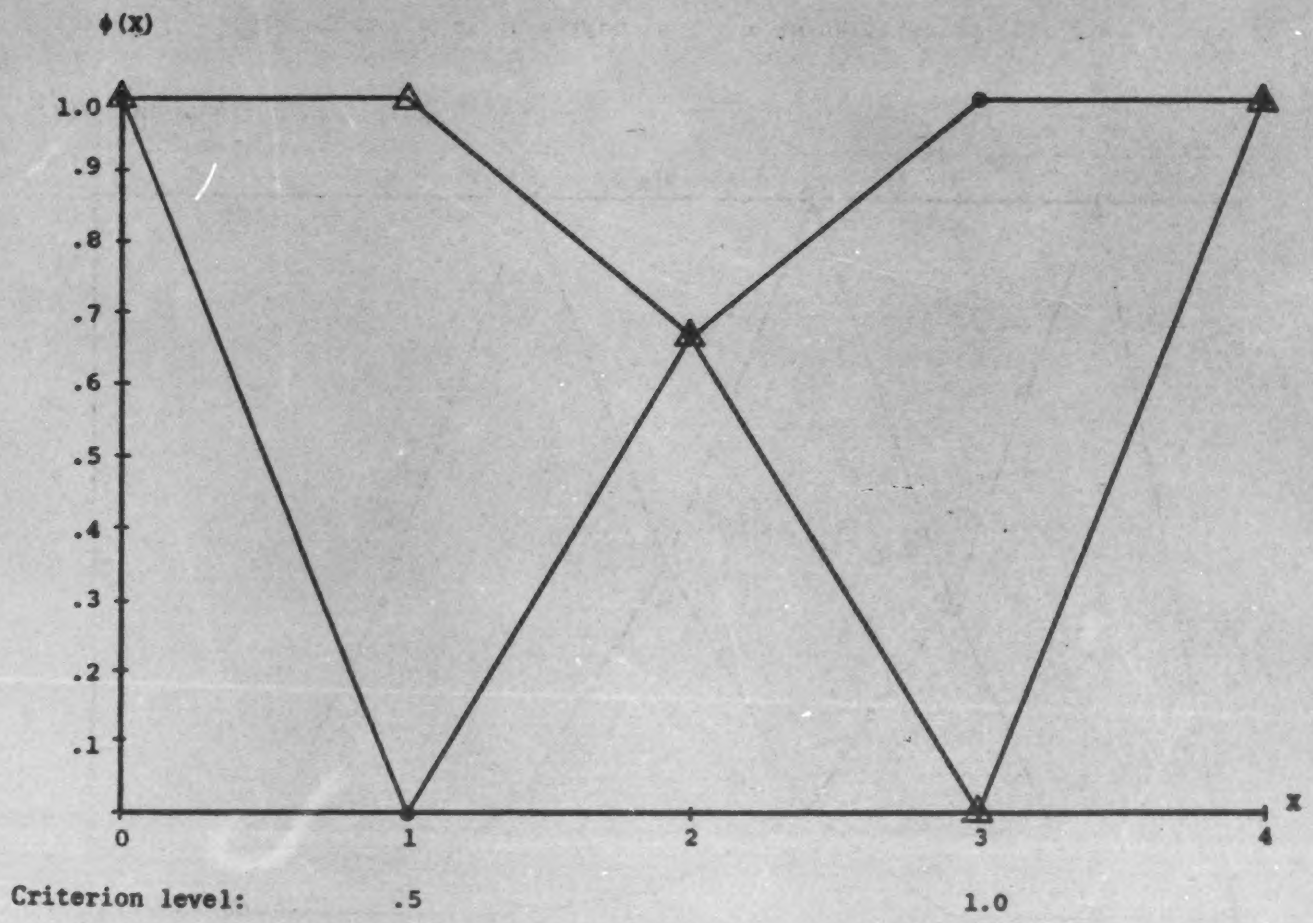


Figure B-1. ϕ(X) for each X on a 4-item test for all meaningful criterion levels.

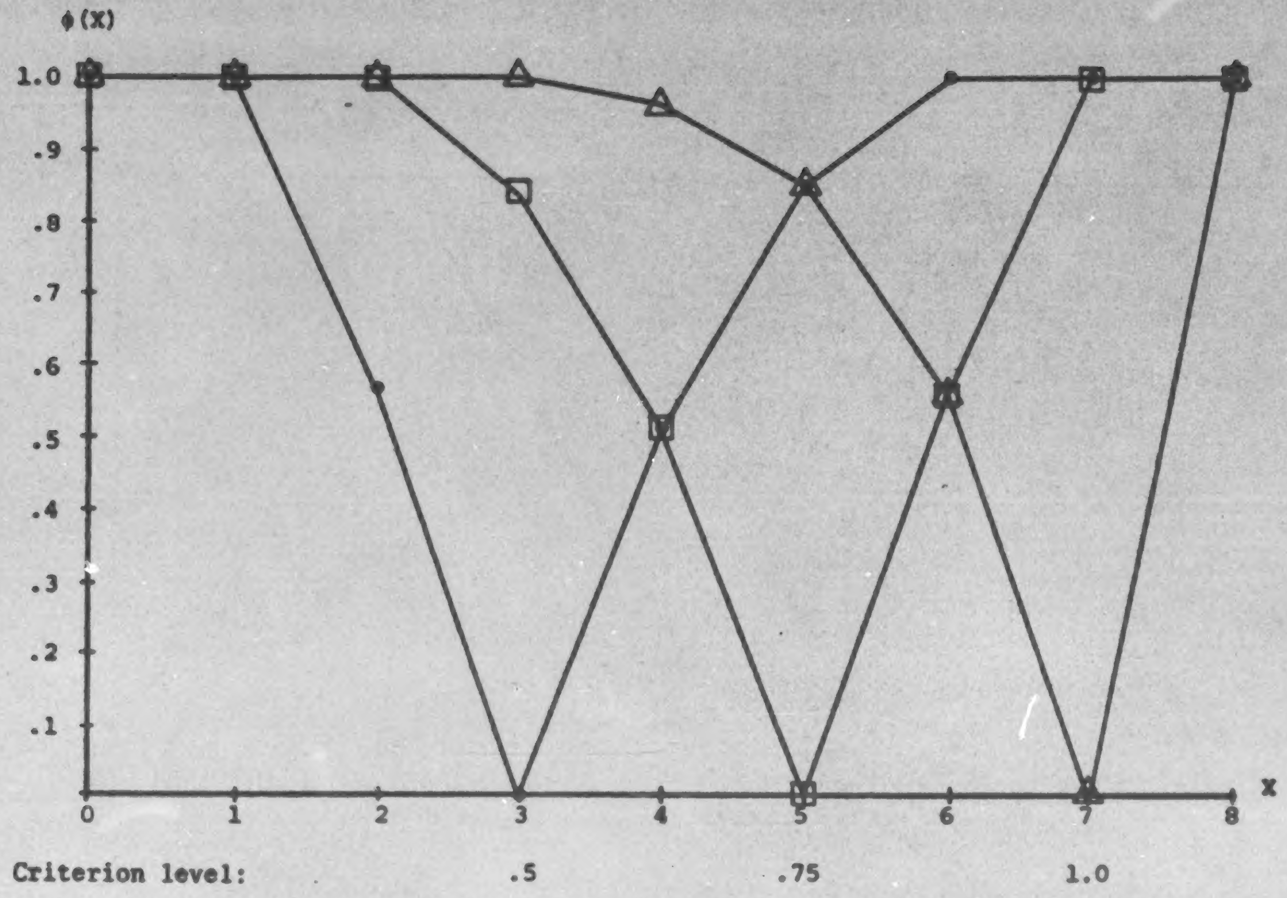


Figure B-2. ϕ(X) for each X on an 8-item test for three selected criterion levels.

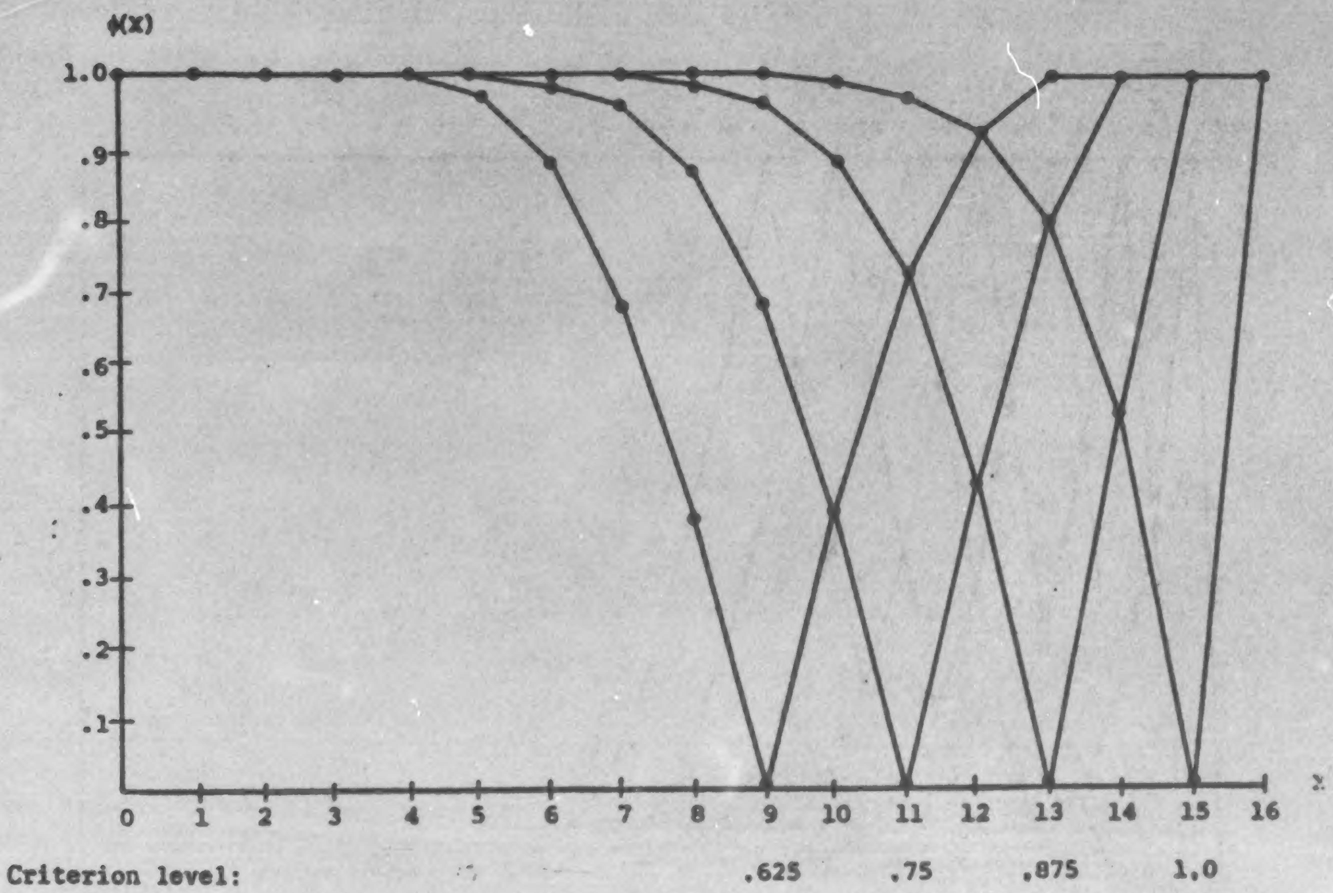


Figure B-3. ϕ(X) for each X on a 16-item test for four selected criterion levels.

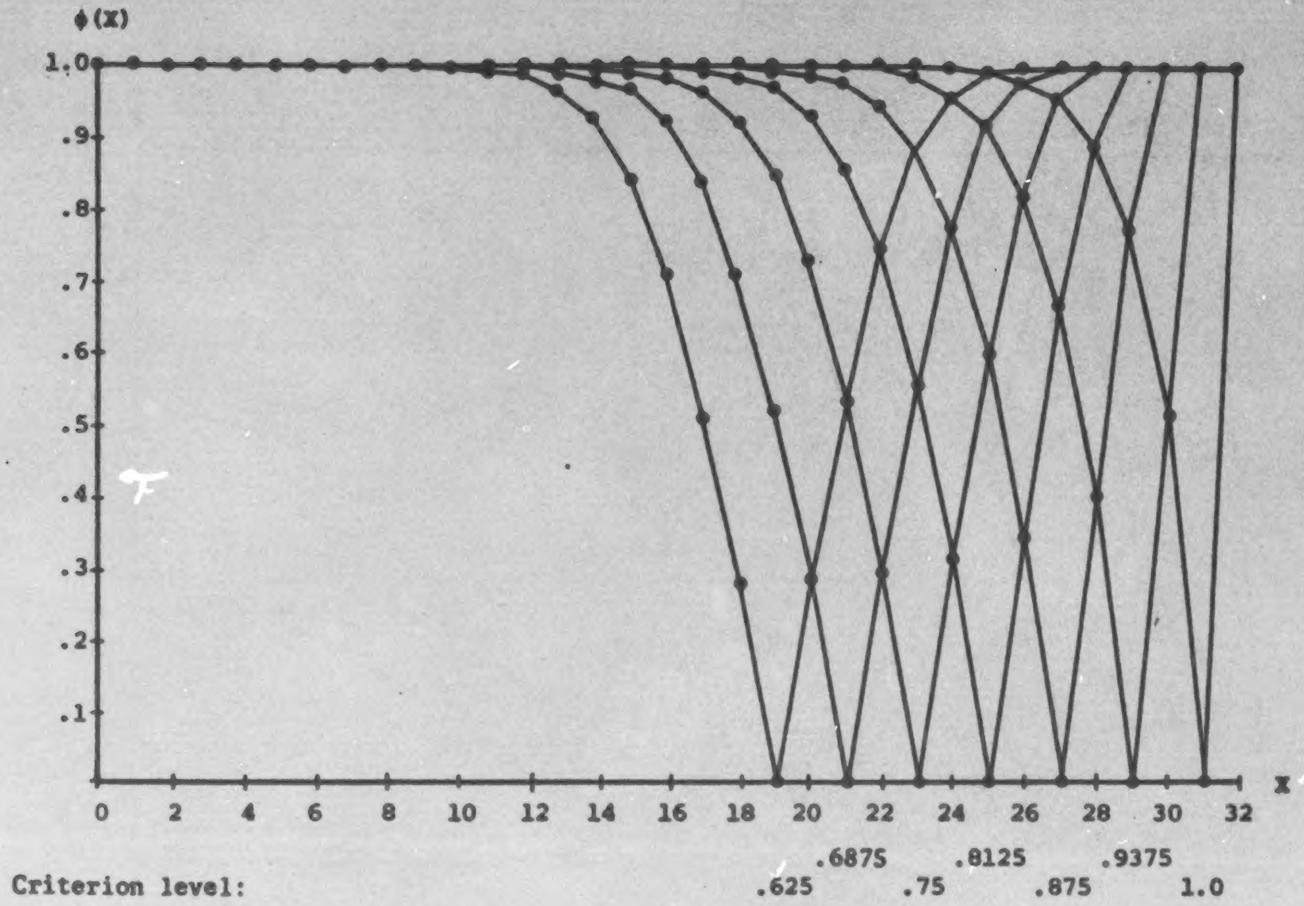


Figure B-4. ϕ(X) for each X on a 32-item test for seven selected criterion levels.

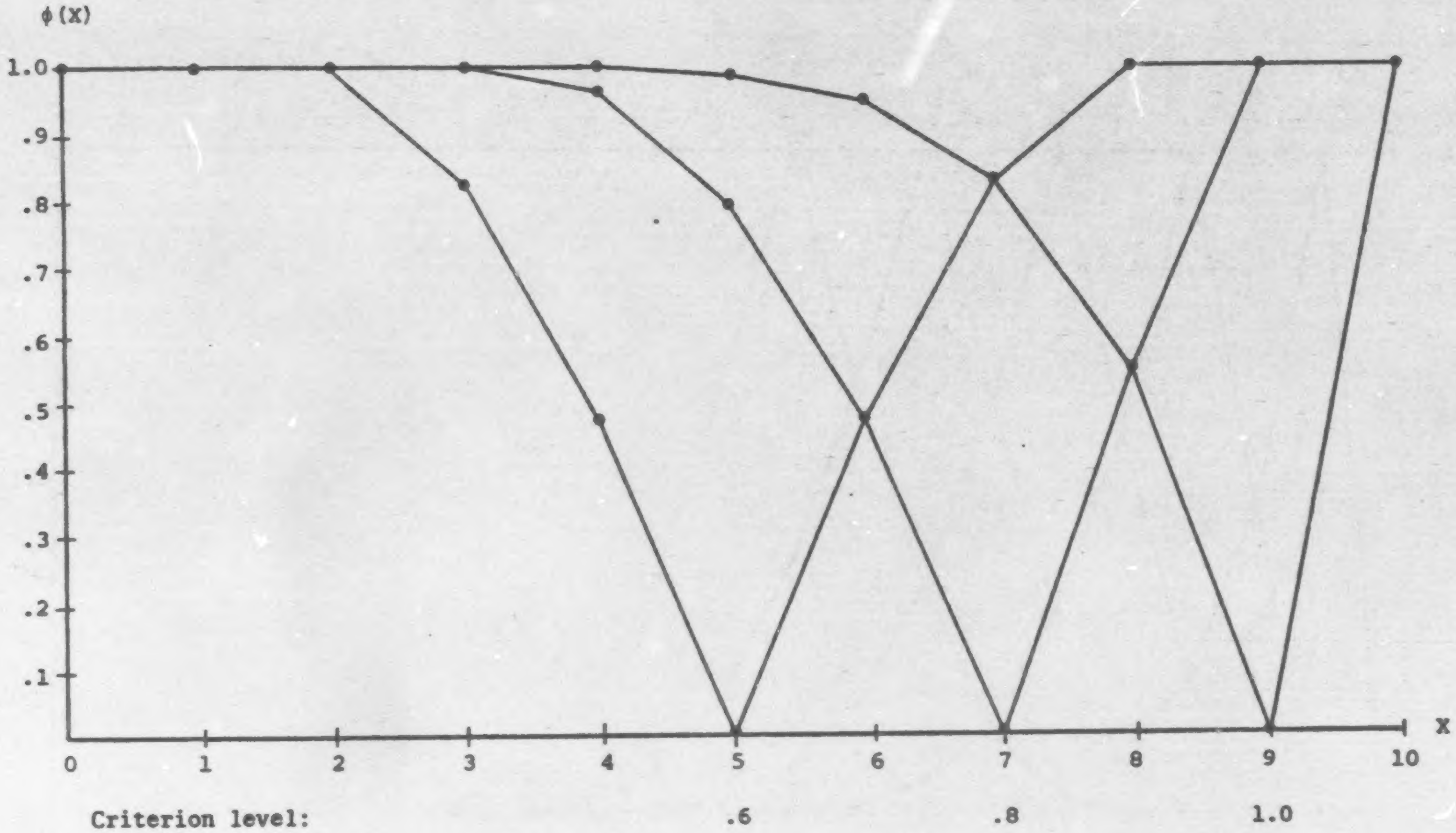


Figure B-5. φ(X) for each X on a 10-item test for three selected criterion levels.

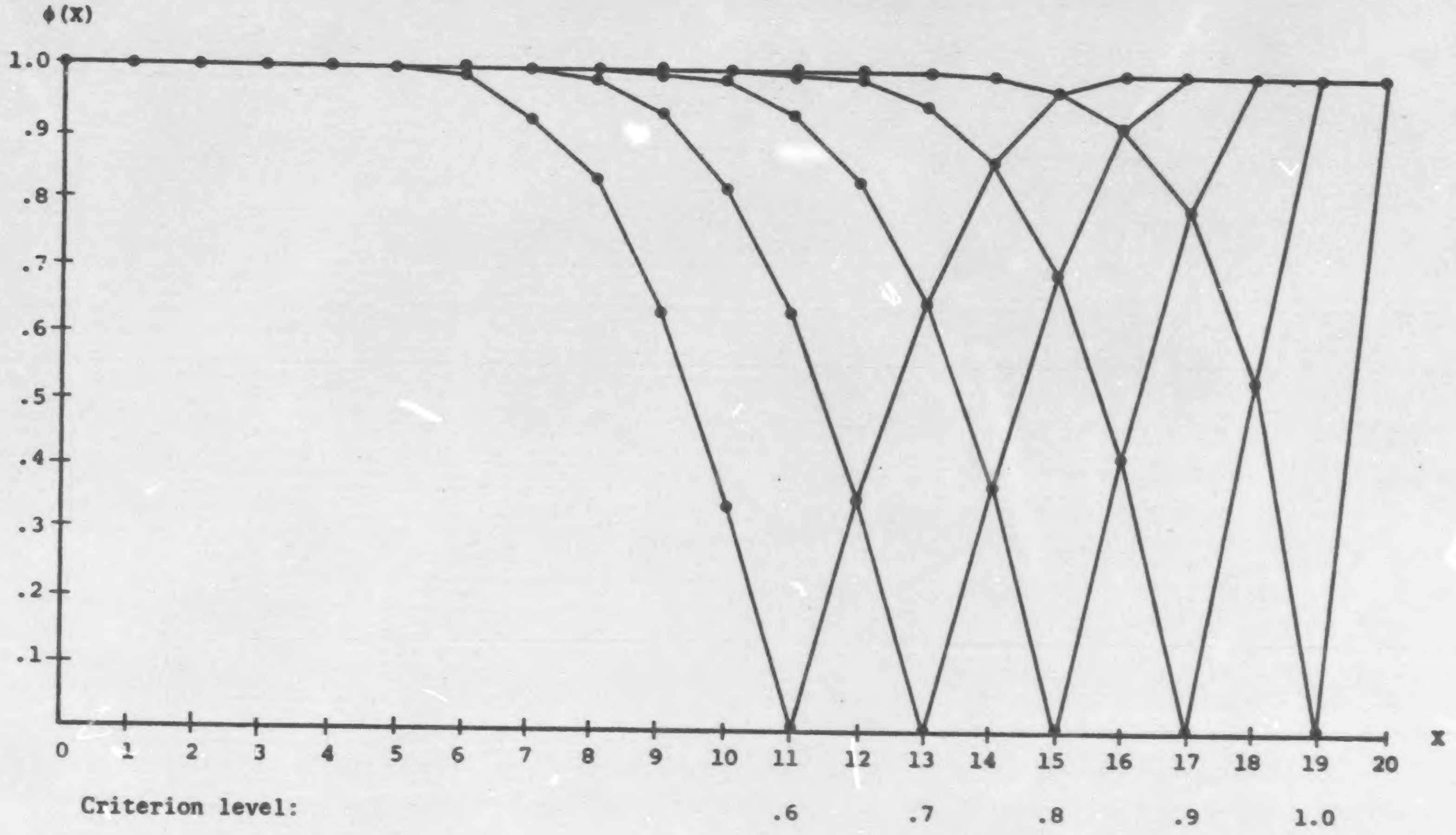


Figure B-6. $\phi(X)$ for each X on a 20-item test for five selected criterion levels.

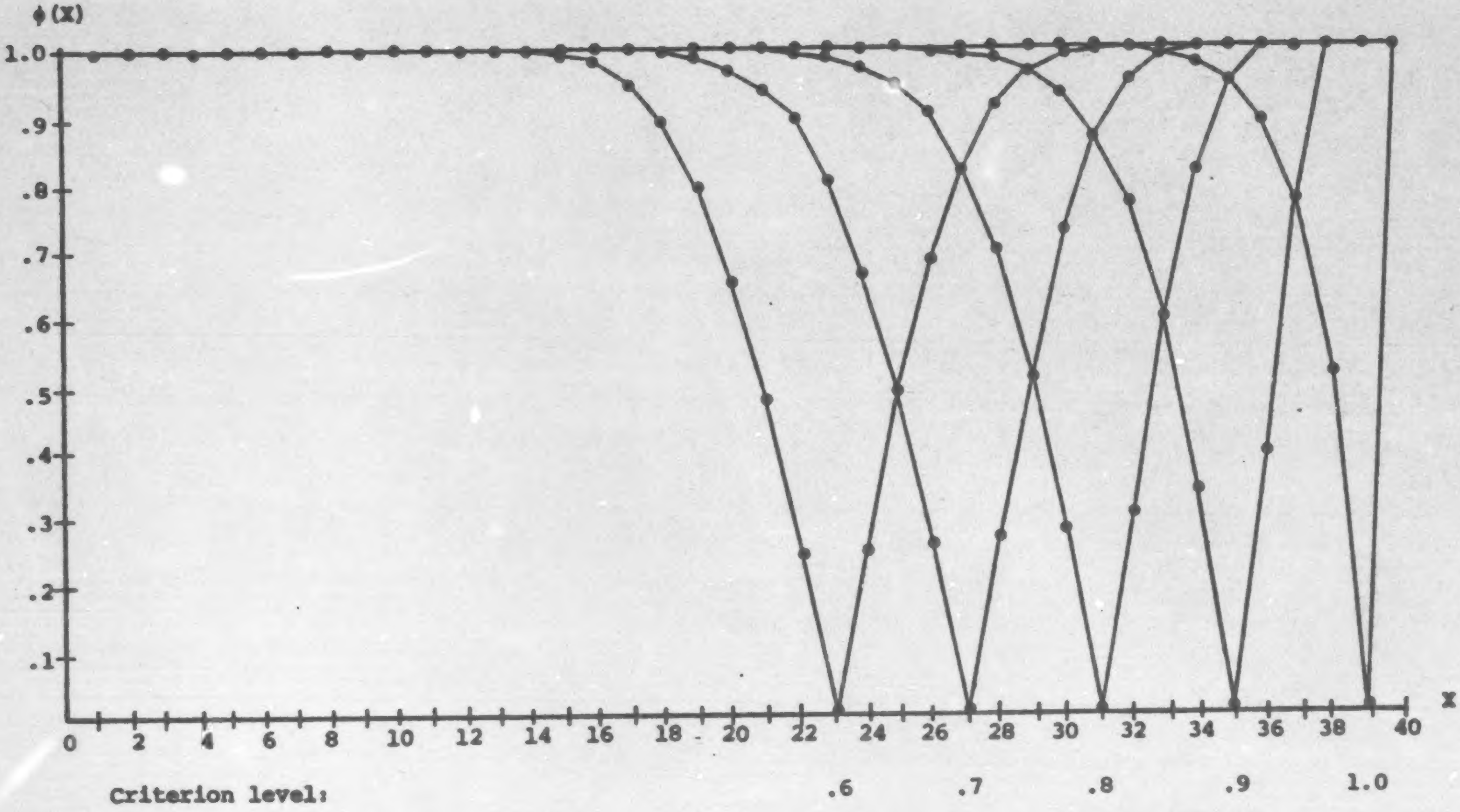


Figure B-7. φ(X) for each X on a 40-item test for five selected cirterion levels.

APPENDIX C

Computer Program Input Parameter Distributions and Subroutines, with Notes on Calculation of Vector Components

Person Competence

$\vec{c} = (c_1, c_2, \ldots, c_p, \ldots, c_N)$, where N = number of persons

1. Chi-square. Calculated from

$$[2^{\nu/2}\,\Gamma(\tfrac{\nu}{2})]^{-1} \int_0^{y_p} t^{(\nu-2)/2} e^{-t/2}\, dt = \frac{p - 0.5}{N}$$

where ν = a parameter selected to control the shape (degrees of freedom)

and $c_p = y_p \cdot A$, where A is a scaling factor chosen so that the maximum value of c_p coincides with a parameter selected to control the range.

2. Mirror-image quasi-chi-square. This is calculated as above, with each c_p being replaced by $1 - c_p$.

The calculation of the chi-square vector components is similar to that of normal distribution vector components (q.v. for a less technical explanation.)

The chi-square distribution was included as an option because empirical data from criterion-referenced tests suggest that post-instruction total score distributions often approximate the distribution of the mirror-image chi-square. Further, it seems reasonable to assume that a population that is not knowledgeable might have a pre-instruction total score distributiion approximating a positively skewed chi-square.

3. Normal. This is calculated from

$$(2\pi)^{-1/2}\int_{-\infty}^{y_p} e^{-t^2/2}\, dt = \frac{p - 0.5}{N}$$

and $c_p = y_p \cdot B + \mu$, where B is a scaling factor to make the
components fit within the predetermined
range, which is itself a parameter selected
to control dispersion

and $\mu = \varepsilon(c_p)$ is a parameter selected to control location.

The vector components are not determined by generating random values, thereby necessitating truncation to make them fit within a range, but rather by apportioning the area under the curve according to the distribution function, and assigning as values the "weighted midpoints" of the line segments within each of N regions. The operation can be thought of as having three steps: first, the mean and standard deviation of the normal distribution are defined; second, the "midpoint" of each segment is found (in the case of the two extreme chunks, by finding the points beyond which in each direction $1/2N$ of the area lies) and third, a linear transformation is applied so that the two extreme values coincide with the limits of the predefined range. (Actually, the range, rather than the standard deviation, is defined, but the computer program merely works backwards.)

4. Bimodal "inverse normal." First a normal distribution vector is generated as defined above. Then a transformation is applied and adjusted. The effect of the transformation and the adjustments is that of cutting the normal distribution in half at the middle, translating the left half .5 to the right, and the right half .5 to the left. (See Figure 6 for an example.)

5. Almode. This is a highly flexible subroutine included to enable

one to approximate unusual shapes in the competence distribution. Given a distribution transcribed into graph form, with x and y coordinates of up to twelve points on the curve such that $0 \leq x_1 < x_2 \ldots < x_n \leq 1$, one inputs these ordered pairs as parameters. The subroutine calculates the areas of the trapezoids under the curve and assigns elements of the competence vector accordingly.

6. List. With this option, one can specify the vector components by supplying a list of the component values.

7. Call. Additional distribution subroutines, such as binomial, can be called into play and used as the need arises. Only options 3 and 4 were used in this study.

Item Difficulty

$\vec{d} = (d_1, d_2, \ldots, d_i, \ldots, d_n)$, where n = number of items

1. "House." This is so named because the region under the curve looks like a child's drawing of a house--an isosceles triangle atop a rectangle. Input parameters define the "corners" and "peak" of the "roof." This distribution includes the degenerate subcases of uniform (rectangular), triangular, and constant.

Empirical data suggest that the distribution of item difficulties often approximates some type of "house" distribution. Uniform distributions were used in this study.

2. Normal. This is the same as described for the vector of person competence.

3. List. This is the came as for $\vec{c}$.

4. Call. This is the same as for $\vec{c}$.

Item Goodness

$$\vec{g} = (g_1, g_2, \ldots, g_i, \ldots, g_n)$$

Distributions and other options for the vector $\vec{g}$ are the same as for $\vec{d}$. However, since the vector components for $\vec{d}$ and $\vec{g}$ are generated in ascending numerical order, a subroutine is employed which randomly permutes the vector components by reassigning their subscripts. This is done in order to avoid interaction between d_i and g_i.

In this study, only uniform distributions were used.

Error Terms

All error terms are randomly generated from an internal normal distribution subroutine, the standard deviation of which can be specified. The starting point (within the computer's subroutine) for any of the error terms can be specified, so that identical error components can be generated on successive trials if this is wanted. This would be desirable, for example, if one wanted to investigate the effect on reliability indices when only the item goodness vector is changed.

APPENDIX D

Summaries of Stepwise Analyses of Regression

D-1. Summary of stepwise analysis of regression, with β as dependent variable.

D-2. Summary of stepwise analysis of regression, with k^2_{TX} as dependent variable.

D-3. Summary of stepwise analysis of regression, with μ^2_c as dependent variable.

D-4. Summary of stepwise analysis of regression, with S_c as dependent variable.

Parameter Set(s)	Coefficient of Determination	Coefficients of regression equation									
		Constant	$\bar{p}$	%v	KR-21	Criterion Level	(a)	k^2_{TX}	(a)	S_c	(a)
1	.91	.29						(b) 1) .72	88	2) .40	
2	.94	.16		(b) 4) -1.4		3) .11		1) .64	90	2) .95	
3	.74	.03						1) .90	71	2) .05	
4	.74	-.68				1) .11		2) 1.56		3) .044	
5	.93	-.06				3) .58		1) .29	83	2) 1.6	
6	.80	11	4) -.57		5) -16.4	2) 1.7	18	3) 5.4	31	1) 6.1	29
7	.97	-.05				3) -.051		1) .88	92	2) .86	
8	.65	-5.1				2) .15		3) .55	10	1) .95	52
unimodal (1,2,3,5,7)	.89	.22	5) -.098	4) -.36	2) -.14			1) .90	83	3) .20	
bimodal (4,6,8)	.20	.85		2) -.19		3) .032				1) .28	15
all (1-8)	.81	.23	5) -.14	4) -.099	2) -.16			1) .91	72	3) .17	

(a) percent of variance in β accounted for by the variable, if $\geq$ 10%
(b) number set off to the left indicates variable's order of entry into regression equation

Table D-1. Summary of stepwise analysis of regression, with β as dependent variable.

Parameter Set(s)	Coefficient of Determination	Coefficients of regression equation										
		Constant	β	(a)	μ^2_c	(a)	S_c	(a)	r_ϕ	(a)	r_{cospi}	(a)
1	.87	-.04	(b) 1) 1.0									
2	.92	.31	1) .70	90	3) -.19						2) .034	
3	.81	.40	1) .84	71							2) -.24	11
4	.72	1.0			1) -.044	69	2) -.012					
5	.95	.95	(c)		1) -.27	92	2) .14				3) .019	
6	.93	.99							1) -.047	93		
7	.95	.74	1) .24	92	3) -.24				2) .14			
8	.83	.99			1) -.018	83						
unimodal (1,2,3,5,7)	.86	.17	1) .81	83	4) -.10				3) .17		2,-5) (d)	
bimodal (4,6,8)	.62	.97			1) -.072	25	2) .037	30	3) .044			
all (1-8)	.80	.41	1) .55	72	4) -.23		6) .055		3) .24		2,-5) (d)	

(a) percent of variance in k^2_{TX} accounted for by this variable, if > 10%

(b) number set off to the left indicates variable's order of entry into regression equation

(c) this is the only unimodal distribution where μ^2_c rather than β is the first entering variable; however, the correlation between μ^2_c and β for this distribution is -.90

(d) r_{cospi} was the second variable to enter, but left at the fifth step

Table D-2. Summary of stepwise analysis of regression, with k^2_{TX} as dependent variable.

Parameter Set(s)	Coefficient of Determination	Coefficients of regression equation									
		Constant	β	(a)	k^2_{TX}	(a)	S_c	(a)	r_ϕ	(a)	r_{cospi}	(a)
1	.90	1.3	(b) 1) -.98	85	3) -.32				2) .32			
2	.93	1.5	1) -.96	88	3) -.59				2) .43			
3	.80	1.5					3) .36		1) 1.9	71	2) -2.5	
4	.88	12.9	3) -8.2	11	1) -6.1	69			2) 4.3		4) -2.1	
5	.96	2.3	4) -.50		1) -2.0	92	3) .33		2) .53			
6	.99	2.5	3) -.85		5) -2.0		2) .76		1) 1.07	97	4) -.053	
7	.99	2.1	1) -1.4	98	4) -.79				2) .49		3) -.040	
8	.99	4.7	3) -1.1		5) -4.3		2) 1.1		1) 1.1	96	4) -.12	
unimodal (1,2,3,5,7)	.91	1.4	2) -1.2	25	5) -.33		3) .24		1) .95	61		
bimodal (4,6,8)	.96	4.9	4) -.48		2) -.47		3) .20		1) .96	94		
all (1-8)	.93	1.4	4) -.81		2) -.71	12	5) .15		1) 1.1	79		

(a) percent of variance in μ^2_c accounted for by this variable, if $\geq$ 10%

(b) number set off to the left indicates variable's order of entry into regression equation

Table D-3. Summary of stepwise analysis of regression, with μ^2_c as dependent variable.

Parameter Set(s)	Coefficient of Determination	Coefficients of regression equation										
		Constant	β	(a)	k^2_{TX}	(a)	μ^2_c	(a)	r_ϕ	(a)	r_{cospi}	(a)
1	.72	-.14	(b) 1) .38	63			1,-4) (c)				2) -.043	10
2	.75	-.08	2) .33				(c)	68			3) -.044	
3	.25	-1.3	1) 2.2	25								
4	(d)	.77										
5	.78	-2.5	1) 1.03	65	4) 1.8		2) .72				3) -.077	
6	.80	-.43	1) .90	29			3) .88	39	2) -1.04	11	4) .041	
7	.90	-.22	1) .45	87							2) -.016	
8	.92	-3.0	1) 1.0	52	5) 2.6		2) .75		3) -.82	23	4) .097	12
unimodal (1,2,3,5,7)	.63	-1.6	3) 1.9		(c)		2) .67					
bimodal (4,6,8)	.56	-11	4) .85		2) 10.4	26	3) .82		1) -.38	18		
all (1-8)	.57	-1.3	2) 1.1				3) .43		(c)			

(a) percent of variance in S_c accounted for by this variable, if $\geq$ 10%

(b) number set off to the left indicates variable's order of entry into regression equation

(c) this was the first variable to enter, but left at the fourth step

(d) no variables entered into this equation, and thus there is no coefficient of determination

Table D-4. Summary of stepwise analysis of regression, with S_c as dependent variable.

APPENDIX E

A Binomial Model for Stepping Up Coefficient Beta

It was noted in Chapter VII that since β is equal to the mean proportion of agreement on all possible split halves of a test, it can be considered to be a half-test coefficient, and thus should somehow be stepped up in order to represent the operational reliability of a whole test. The formula presented in Chapter 6 was based on purely empirical evidence, and thus is unsatisfying mathematically.

One mathematical approach to the solution to this problem is to use the binomial probability model. Briefly, the method is to calculate β from an estimated frequency distribution of total scores for a double-length (2n items) test, based on the obtained frequency distribution of scores from the test of n items, and utilizing the binomial probability model to estimate likelihoods concerning each person's double-length test score.

More specifically, suppose person p receives a score of x on an n-item test. Under the binomial model, $\frac{x}{n}$ is the best estimate of the proportion of items in the universe that he would answer correctly, and hence also the best estimate of the proportion of items he would answer correctly on a test of 2n items. Let Y_p be the examinee's score on this test; $Y_p \varepsilon \{0, 1. \ldots, 2n\}$. Then the probability that person p receives a score of y, i.e.,

$$Pr(Y_p = y \mid X_p = x), \text{ is } \binom{2n}{y}\left(\frac{x}{n}\right)^y\left(\frac{n-x}{n}\right)^{2n-y}.$$

(Note here that

$$\sum_{y=0}^{2n} Pr(Y_p = y \mid X_p = x) = \sum_{y=0}^{2n} \binom{2n}{y}\left(\frac{x}{n}\right)^{y}\left(\frac{n-x}{n}\right)^{2n-y} = 1.)$$

But there are f_x persons with score x, and hence the contribution to y from all those with this score is

$$f_x \binom{2n}{y}\left(\frac{x}{n}\right)^{y}\left(\frac{n-x}{n}\right)^{2n-y}.$$

However, a number of different scores x will contribute to the frequency of y. Thus, summing over all scores x, the frequency of score y in the distribution is

$$F_y = \sum_{x=0}^{n} f_x \binom{2n}{y}\left(\frac{x}{n}\right)^{y}\left(\frac{n-x}{n}\right)^{2n-y}.$$

We have thus arrived at a method of calculating expected frequencies of each component of the vector $\overrightarrow{F_y} = (F_0, F_1, \ldots, F_{2n})$, the expected frequency distribution of scores on the (hypothetical) double-length test. We can now compute β on the double-length test:

$$\beta = \frac{1}{N}\sum_{y=0}^{2n} F_y \Phi_y$$

$$= \frac{1}{N}\left(\sum_{y=0}^{C-1} F_y + \sum_{y=C}^{2C-2} F_y \, \Phi_y(y-[C-1], C-1) + \sum_{y=2C}^{n+C-1} F_y \, \Phi_y(C, y-C) + \sum_{y=n+C}^{2n} F_y\right)$$

where (as before)

N = number of examinees;

n = number of items (on the single-length test);

and

y = a score on the hypothetical test of 2n items;

C = the cutoff score on the n-item test, and hence the smallest integer $\geq$ cn;

$$\phi_y(a,b) = \sum_{j=a}^{b} \frac{\binom{y}{j}\binom{2n-y}{n-j}}{\binom{2n}{n}}$$ and

$$F_y = \sum_{x=0}^{n} f_x \binom{2n}{y}\left(\frac{x}{n}\right)^y\left(\frac{n-x}{n}\right)^{2n-y}$$, where f_x is the obtained frequency of score X on the n-item test.

Note that

1. F_y is generally not an integer;
2. when x = 0 or x = n, the quantity 0^0 appears in the formulation of F_y, and must be defined as equal to 1;
3. the second term in the brackets vanishes when C = 1; the third term vanishes when C = n;
4. the adjustment for odd n is no longer necessary;
5. an analogous formula holds for β_3, the stepped-up coefficient for trichotomous data.

National Evaluation Committee

Francis S. Chase, Chairman
Emeritus Professor
University of Chicago

Helen Bain
Past President
National Education Association

Lyle Bourne
Professor
University of Colorado

Sue Buel
Consultant, Portland, Oregon

Roald F. Campbell
Emeritus Professor
The Ohio State University

George E. Dickson
Dean, College of Education
University of Toledo

Larry R. Goulet
Professor
University of Illinois

Chester W. Harris
Professor
University of California - Santa Barbara

William G. Katzenmeyer
Professor
Duke University

Barbara Thompson
Superintendent of Public Instruction
State of Wisconsin

Joanna Williams
Professor
Teachers College
Columbia University

University Advisory Committee

John R. Palmer, Chairman
Dean
School of Education

William R. Bush
Deputy Director
R & D Center

David E. Cronon
Dean
College of Letters and Science

Diane H. Eich
Specialist
R & D Center

Evelyn L. Hoekenga
Coordinator
R & D Center

Dale D. Johnson
Associate Professor
Curriculum and Instruction

Herbert J. Klausmeier
Member of the Associated Faculty
R & D Center

James M. Lipham
Member of the Associated Faculty
R & D Center

Wayne R. Otto
Associate Director
R & D Center

Richard A. Rossmiller
Director
R & D Center

Elizabeth J. Simpson
Dean
School of Family Resources
and Consumer Sciences

Len Van Ess
Associate Vice Chancellor
University of Wisconsin - Madison

Associated Faculty

Vernon L. Allen
Professor
Psychology

B. Dean Bowles
Professor
Educational Administration

Thomas P. Carpenter
Assistant Professor
Curriculum and Instruction

Marvin J. Fruth
Professor
Educational Administration

John G. Harvey
Professor
Mathematics
Curriculum and Instruction

Frank H. Hooper
Professor
Child Development

Herbert J. Klausmeier
V.A.C. Henmon Professor
Educational Psychology

Joseph T. Lawton
Assistant Professor
Educational Psychology

Joel R. Levin
Professor
Educational Psychology

L. Joseph Lins
Professor
Institutional Studies

James M. Lipham
Professor
Educational Administration

Donald N. McIsaac
Professor
Educational Administration

Gerald Nadler
Professor
Industrial Engineering

Wayne R. Otto
Professor
Curriculum and Instruction

Robert G. Petzold
Professor
Music
Curriculum and Instruction

Thomas S. Popkewitz
Assistant Professor
Curriculum and Instruction

Thomas A. Romberg
Professor
Curriculum and Instruction

Richard A. Rossmiller
Professor
Educational Administration

Dennis W. Spuck
Assistant Professor
Educational Administration

Michael J. Subkoviak
Assistant Professor
Educational Psychology

Richard L. Venezky
Professor
Computer Sciences

J. Fred Weaver
Professor
Curriculum and Instruction

Larry M. Wilder
Assistant Professor
Child Development